Inequality, inequality of opportunity, and growth: what are we talking about? Theory and empirical investigation in Brazil *

Geoffrey Teyssier
(June 2013)

Université Paris 1 Panthéon-Sorbonne
UFR 02 Sciences Economiques
M2R Globalization (option Development Economics)
Supervisor: Charlotte Guénard
Co-supervisor: Sandra Poncet

Building on the existing literature, a synthetic approach intended to ease the understanding of the notion of inequality of opportunity is developed. In turn, this paper tests a convincing hypothesis explaining the mixed evidence found by empirical studies regarding the instrumental effect of inequality on growth: income inequality would in fact be a composite measure of inequality of opportunity, which is expected to be detrimental to growth, and of inequality effort, which is expected to be beneficial; the effect of total income inequality would then depend on which sort of inequality dominates. This hypothesis, already confirmed by Marrero and Rodríguez (2012) in the US, needs to be validated in other countries and on different samples in order to gain legitimacy. This paper consequently replicates the benchmark regressions from Marrero and Rodríguez (2012) in an emerging economy, namely Brazil. The results are in complete contradiction with those found in the US: neither inequality of opportunity nor inequality of effort have a significant impact on growth, whatever the econometric specification used.

JEL Classification Codes: D63, E24, O15, O40.

Key Words: income inequality; inequality of opportunity; economic growth

* I thank Charlotte Guénard for her guidance and support, as well as Sandra Poncet, Denis Cogneau and Francisco H.G. Ferreira for helpful comments. Lastly, I would also like to thank my family and friends for giving me the “opportunity” to share my thoughts with them on this challenging topic.

"Accipere quam facere praestat injuriam."
It is better to suffer an injustice than to commit one.
(Cicero)

I. Introduction

Do notions of equity and equality matter when it comes to economic performance, and more precisely, when it comes to economic growth? Economists have tried to answer this question mainly by studying the causal impact going from inequality to growth. Until the mid-1990s, cross-section studies seemed to reach the consensus that initial inequality was detrimental to growth (Bourguignon, 1993; Alesina and Rodrik, 1994; Persson and Tabellini, 1994; Birdsall, Ross and Sabbot, 1995; Clarke, 1995; Perotti, 1996). These empirical studies were supported by a good deal of theoretical models explaining how inequality might hamper development.[1] However, the consensus broke down after the introduction of Deininger and Squire's (1996) dataset. This dataset on income inequality has been an important milestone in the growth-inequality literature: studies based on this "high quality" dataset were less prone to measurement errors, and perhaps more importantly, they could investigate the inter-temporal dimension of the growth-inequality relationship. After the introduction of this dataset, some authors (Li and Zou, 1998; Forbes, 2000) have posited for a beneficial effect of inequality on growth, while others (Deninger and Squire, 1998; Easterly, 2007) have maintained that it was detrimental.[2] To further complicate things, Barro (2000) states that inequality is detrimental in poor countries and beneficial in rich countries, while even others (Cogneau and Guénard, 2002; Banerjee and Duflo, 2003; Bleaney and Nishiyama, 2004) suggest that inequality does not have any robust significant effect on growth. As summarized by Herzer and Vollmer (2012, p.490), the growth-inequality relationship is controversial, to put it mildly: "if there is anything we can take away from the existing literature […], it is the fact that there is no consensus on the question of whether inequality affects growth positively, negatively, or at all".[3]

A possible explanation for this overall confusion is that income inequality might not be the relevant concept to consider from both a normative and a positive point of view. From

---

[1] See Galor and Zeira (1993) and Piketty (1997) for credit market imperfections models; Bertola (1993), and Alesina and Rodrik (1994) for political economy models; Alesina and Perotti (1996) for a property rights model; or Dahan and Tsiddon (1998) and Galor and Zang (1997) for endogenous fertility models.

[2] Empirical studies positing for a beneficial effect of inequality were supported by various theoretical models too. The main argument is that high inequality can lead to higher amounts of saving and investment due to an increasing marginal propensity to save, in a Kaldorian investment model. Arguments based on political economy models (Saint Paul and Verdier, 1993) and human complementarities (Benabou, 1996), among many others, can also explain how high inequality might, in some cases, prompt growth.

[3] For recent surveys of this literature, see Yusuf (2005) and Ehrhart (2009).

a normative point of view, what the concept of income (in)equality says is that everyone should earn the same income. Intuitively, there is something going wrong with this statement because not everyone is equally deserving. Despite some important differences, theorists of social justice such as Dworkin (1981), Sen (1985), Arneson (1989) or Cohen (1989) share this common idea that those who get a piece of the pie should deserve it. More formally, they put the emphasis on the notion of individual responsibility for assessing whether an income distribution is fair: inequality is inequitable only to the extent that it results from factors outside the sphere of the individuals' control. This is the concept inequality of opportunity, which is morally unacceptable (as opposed to income inequality). In their seminal papers, Roemer (1993, 1998), and van de Gaer (1993) have integrated this concept of inequality of opportunity within a formal theoretical framework. Incomes achieved by individuals are the result of two types of factors: those factors over which they have no control and for which they should not be held responsible, such as family background, race or gender; and those factors that they can control and for which they should be held responsible, such as their length of study or their involvement into work. Following Roemer's terminology (1998), factors of the former kind are called of "circumstances" while factors of the latter kind are called "efforts".[4] Income inequality resulting from different circumstances across individuals corresponds to the inequality of opportunity (IO); it is viewed as morally unacceptable and should be suppressed according to the "compensation principle". On the contrary, income inequality resulting from different levels of effort across individuals corresponds to the inequality of effort (IE); it is morally acceptable and should be left untouched according to "reward principle".

From a positive point of view, it has also been suggested that IO might be a more relevant concept than income inequality for understanding whether more unequal societies are experiencing lower economic performance (World Bank, 2006; Bourguignon et al., 2007b). Because income inequality is a composite measure of IO and IE, which are expected to have contradictory effects on growth, income inequality would be positively or negatively correlated with growth depending on which type of inequality dominates. Inequality would in fact be like cholesterol: we should distinguish the "good" from the "bad" one. Marrero and Rodríguez (2012a) test this "inequality as cholesterol hypothesis". By considering separately IO and IE in a growth regression, they bridge the gap between the macroeconomic growth-

[4] One should be very cautious with the term "effort", which is often defined negatively as all those factors affecting individual income that are not considered as a circumstance. Some authors prefer to it the more neutral term "responsibility factor".

inequality literature and the microeconomic IO literature. Using household survey data in the US, they find the expected results: IO and IE have a robust and significant effect on growth in US states (negative for IO, positive for IE) while the effect of overall inequality is not robust.

This paper seeks to investigate both the normative (ie. moral) and positive (ie. instrumental effect on growth) aspects of IO. Its research question is consequently twofold. The first research question is normative and theoretical in the sense that it has to do with the very concept of IO: how to define and how to measure it? Although the basic idea underlying the theory of IO – that one's achievements should reflect one's own merit rather than one's undeserved and relatively (dis)advantageous set of opportunities – is fairly intuitive, things get more complicated when it comes to measure IO or to define it formally. Much papers have answered this question but there remains a lot to be said about it because the IO literature "is still in its infancy and different, often conflicting, approaches have been proposed" (Checchi and Peragine, 2010). This paper builds on two previous sets of studies – Bourguignon et al (2007a) and Ferreira and Gignoux (2011) on one side; Checchi and Peragine (2010) on the other side – about IO in order to provide a synthetic approach to its definition and measurement. The second research question is positive and empirical. It asks whether the results from Marrero and Rodríguez (2012a) – the only paper having studied empirically the effect of IO on growth so far, to the best of my knowledge – can be generalized to other countries than the US. To this end, an independent of replication of their empirical study is implemented taking as a case study an emerging economy, namely Brazil. Results are disappointing: neither IO nor IE have a significant impact on growth in Brazilian states during the 1980-2009 period.

The rest of this paper is organized as follows. Section 2 seeks to answer the first question and is about the concepts and measurements of IO. The synthetic approach developed in this section also helps to derive some theoretical conditions under which IO *might* be detrimental or beneficial to growth. The empirical part of this paper – devoted to the second research question – starts with section 3, which describes the micro-data samples used to compute measures of IO and IE in Brazil. In particular, a procedure of multiple imputation is developed in this section because these micro-data samples are suffering from sample bias due to missing observations on parental circumstances. Section 4 answers the empirical research question by providing an independent replication of the benchmark regressions from Marrero and Rodríguez (2012a), taking Brazilian states as the units of analysis. Lastly, section 5 concludes.

II. Concepts and measurements of inequality of opportunity

Although there has been much debate in the philosophical literature about where to draw the line between circumstances and efforts, the basic idea underlying the reward and compensation principles is quite intuitive: income differences across individuals are fair to the extent that they are the result of factors that they can control (ie. they are the result of their own individual decisions). Things get more complicated when turning to devise a measure of IO, for two main reasons. The first obvious reason is that the full set of circumstances and efforts affecting individual income is never observed in practice. The second reason is that we need to define precisely the concept of *equality* of opportunity (EO) in order to understand what IO, *a contrario*, refers to. This is not as straightforward as it may seem. There exist two main definitions of EO in the literature. The "ex-ante" one, from van de Gaer, seeks to reduce income differences "between groups of people who share the same circumstances". The "ex-post" one, from Roemer, seeks to reduce income differences "among people who have exerted the same degree of effort, regardless of circumstances" (Ferreira and Gignoux, 2011).[5] The distinction between these two definitions is tricky: they embody the same idea but Fleurbaey and Peragine (2013) have proven that they are in fact incompatible with each other.

The rest of this section provides a non-exhaustive overview of the concept IO. It identifies two prominent approaches to the measurement of IO. The first approach is from Checchi and Peragine (2010) and relies on a nonparametric decomposition of inequality into its within and between components. This approach is particularly helpful for the purpose of understanding the difference between ex-ante and ex-post IO. The second approach is from Bourguignon et al. (2007a) and Ferreira and Gignoux (2011). It consists in representing individual income as a function of circumstances and efforts in order to estimate parametrically a counterfactual income distribution where the sole effect of circumstances remains (and hence where all income differences correspond to IO). A synthesis of these two approaches, which are not exactly equivalent in terms of measurement of IO, is made in a

[5] The ex-ante and ex-post views are the mainstream approaches to IO in the literature but they are not the only ones. For instance, a few papers have also adopted a "reward" view that seeks not to suppress fair income differences. The ex-ante and ex-post views are, on the contrary, said to be closer in spirit from the compensation principle in the sense that they primarily seek to suppress unfair income differences. Pignataro (2012) and Ramos and van de Gaer (2012) provide excellent surveys of these different approaches to IO. For papers focused on the ex-ante and ex-post approaches mainly, see among others Ruiz-Castillo (2003), Bourguignon et al (2007a), Ooghe et al. (2007), Checchi and Peragine (2010), Ferreira and Gignoux (2011), and Brunori et al. (2013). The latter provides a "non-technical" overview of the ex-ante and ex-post approaches that might be highly valuable for the non-specialists about IO.

third step. This synthesis then allows me to derive the conditions under which IO *might* be beneficial or detrimental to growth and, in turn, to observe which conditions Brazil is fulfilling. Lastly, let me point out that this theoretical section about the concepts and measurements of IO is not a matter of pure conceptual refinement: one has to think about the meaning of IO *per se* if one is to understand how IO might impact growth.

1. *The approach from Checchi and Peragine (2010) and the ex-ante/ex-post distinction*

The approach from Checchi and Peragine (2010) implies partitioning the population into groups of people sharing similar characteristics. According to the way the partition is done, one will derive either an ex-ante measure of IO, either an ex-post one. In the ex-ante approach,[6] the population is partitioned into groups called "types" that gather people sharing the same circumstances. It sees IO as the *inequality between types*. To define this *inequality between types*, it is first necessary to define the outcome variable that each type is enjoying from. A natural candidate for this outcome variable is the mean income of each type, which can be seen as the value of the opportunity set faced by its individuals. Ex-ante EO is then achieved when all types have the same mean income, which is another way of saying that the value of the opportunity set is the same for all individuals in the population. Indexing types by the superscript $k \in \{1, ..., K\}$, Checchi and Peragine (2010) show that ex-ante IO can be computed as follows by using the additive decomposability and path-independence properties of the Theil L index, denoted by $E(0)$:[7]

$$(1) \qquad \underbrace{E(0)}_{\substack{total\\inequality}} = \underbrace{\sum_{k=1}^{K} \frac{N_k}{N} \log\left(\frac{N_k / N}{Y_k / Y} \right)}_{\substack{IO \\ (ie.\,inequality\ between\ types)}} + \underbrace{\sum_{k=1}^{K} \frac{N_k}{N} E(0)_k}_{\substack{IE \\ (ie.\,inequality\ within\ types)}}$$

with $E(0)_k$ the inequality within one particular type k, N the number of individuals in the population, N_k the number of individuals in type k, Y_k the total income of type k, Y the

[6] The term "approach" is used alternatively to denote the framework Checchi and Peragine (2010), as opposed to the one from Bourguignon et al. (2007a) and Ferreira and Gignoux (2011); either to denote the ex-ante view of IO, as opposed to the ex-post one.

[7] The Theil L index shares with most inequality measures the four basic properties of anonymity, scale independence, population independence, and transfer principle. It shares with any other members of the Generalized entropy class the fifth property of additive decomposability (Bourguignon, 1979; Shorrocks, 1980; Cowell, 1980; Foster, 1985), meaning that it can be split into a within and between component for any arbitrary partition of the population. It is the only index to possess additionally the sixth property of path-independence (Foster and Shneyerov, 2000), meaning that its within and between components exactly sum up to total inequality. This sixth property makes it the preferred index in the literature for the purpose of measuring IO. See Ferreira and Gignoux (2011) for more details about the desirable properties of the Theil L index in the context of measuring IO.

aggregate income of the population, and K the total number of types. Clearly, since all individuals from the same type share the same circumstances, IE is the inequality within types (where income differences are due to efforts) and IO is the inequality between types (where income differences are due to circumstances).

While the ex-ante approach focuses on unfair income differences between people with different circumstances, the ex-post approach focuses on unfair income differences between people having exerted the same level of effort. It consists in partitioning the population into groups called "tranches" that gather people having exerted the same effort level.[8] Ex-post IO is then the inequality within tranches since income differences within a tranche can only be attributable to different circumstances (and ex-post IE is the inequality between tranches). However, a difficulty arises here because the effort level provided by an individual is correlated with his circumstances: a typical example is the fact that children from well-educated parents (a circumstance) tend to follow longer studies (an effort) than others. Roemer (1993, 1998) posits that these differences in efforts due to circumstances be treated as characteristics of the type rather than of the individual (and hence, as circumstances effectively). He suggests a proxy for the effort level not explained by circumstances: the income quantile within each type the individual belongs to. Since income is monotonically increasing in efforts, splitting individuals within each type according to their income quantile will reflect their effort level. In this framework, a tranche thus gathers individuals belonging to the same income quantile across types. This is known as the "quantile hypothesis". Whether one adopts this hypothesis or not, defining ex-post EO as a situation where "all individuals having exerted the same effort earn the same income" – as is often the case in the literature – is in any case misleading. The concepts of *absolute* and *relative* efforts are helpful for the purpose of defining ex-post EO without ambiguity.[9] Absolute effort corresponds to the commonly understood and concrete notion of effort (such as one's length of study). Relative effort corresponds to what Roemer had in mind and is an abstract notion. It is the effort level exerted by an individual *relatively to* his type. "Relatively to" is not a synonym for "conditionally on" here. Relative effort is by definition an idiosyncratic characteristic: it is unconditional on one's type for the very reason that it is measured in comparison with the effort levels exerted by the other members of the corresponding type. With this distinction in mind, ex-post EO can be unambiguously defined as a situation where "all individuals having

[8] The term "tranche" was coined by Peragine (2004). The term "type" was coined by Roemer (1993).

[9] To the best of my knowledge, the concepts of absolute and relative efforts have been introduced explicitly by Nilsson (2005). They have, unfortunately, seldom been used since then.

exerted the same level of *relative effort* earn the same income".

Checchi and Peragine (2010) have shown that both ex-ante and ex-post IO can be computed by applying the Theil L index over a counterfactual income distribution, instead of applying it over the actual income distribution and decomposing it into its within and between components (as was done in equation (1)). As a matter of fact, the two methods – decomposition of the actual distribution versus computation of a counterfactual distribution – are exactly equivalent.[10] For instance in the ex-ante approach, replacing the income Y_i^k of each individual *i* belonging to type *k* by the mean income μ^k of his type yields the "smoothed income distribution" $\{\mu_i^k\}$ where all IE has been suppressed (because all individuals belonging to the same type have the same income in this counterfactual distribution) and where, as a consequence, only IO remains. Alternatively, replacing the income Y_i^k of each individual by $Y_i^k \frac{\mu}{\mu^k}$, where μ denotes the mean income of the overall population, yields the "standardized income distribution" $\{v_i^k\}$ where all IO has been suppressed (because the income of each individual is rescaled such that all types have the same mean income in this counterfactual distribution) and where, as a consequence, only IE remains.[11]

2. *The approach from Bourguignon et al. (2007) and Ferreira and Gignoux (2011)*

The theory of IO identifies circumstances and efforts as the main determinants of individual income. Individual income can thus be represented as a function of observed circumstances C_i^k, observed efforts E_i^k and unobserved determinants u_i^k:

(2) $$Y_i^k = f(C_i^k, E_i^k, u_i^k),$$

where C_i^k is a vector of circumstances variables, and where E_i^k is a vector of effort variables. Bourguignon et al. (2007a) notice that circumstances C_i^k are economically

[10] To see that the two methods are equivalent in the case of the ex-ante smoothed income distribution, notice that applying the Theil L index without decomposition over this counterfactual distribution means replacing the term Y_i by μ_i^k in the following formula (which corresponds to the Theil L index not decomposed and applied over the actual income distribution):

$$E(0) = \sum_{i=1}^{N} \frac{N_i}{N} \log\left(\frac{N_i/N}{Y_i/Y}\right)$$

[11] Notice that the opposite holds true regarding the ex-post approach: replacing the income Y_i^T of each individual i belonging to tranche T by the mean income μ^T of his tranche yields a smoothed distribution where only IE remains; while replacing Y_i^T by $Y_i^T \frac{\mu}{\mu^T}$ yields a standardized distribution where only IO remains.

exogenous – in the sense that they cannot be influenced by any individual decision – but that efforts E_i^k are influenced by circumstances as well as unobserved determinants *v*.[12] Hence, circumstances have two kinds of effect on individual income: a direct one, and an indirect one *through* efforts. Bourguignon et al. (2007a) make a serious breakthrough in the literature by suggesting to rewrite equation (2) more fully as:

(3) $$Y_i^k = f(C_i^k, E_i^k(C_i^k, v), u_i^k)$$

which can in practice be estimated by the following system of structural equations:

(4a) $$Y_i^k = C_i^k \alpha + E_i^k \beta + u_i^k$$

(5a) $$E_i^k = HC_i^k + v$$

where α and β are vectors of coefficients, and *H* is a matrix of coefficients (because the dependent variable E_i^k in equation (5a) is treated as a vector). The problem with estimating equations (4a) and (5a) is that the full set of effort and circumstance variables is never observed in practice. OLS estimates of α, β and *H* would thus suffer from omitted variable bias (and possibly from reverse causality and measurement error as well). To deal with the endogeneity of the regressors C_i^k and E_i^k, an instrumental variable approach is unlikely to succeed: it is almost impossible to find a true source of exogenous variation in C_i^k or E_i^k that would be uncorrelated with all unobserved determinants of income and efforts included respectively in the error terms u_i^k and *v*.

Bourguignon et al. (2007a) show that, fortunately, this problem of endogeneity is not a real problem because we are only interested in measuring IO, not in estimating a causal relationship between C_i^k, E_i^k and Y_i^k. Hence, instead of estimating the structural model (4a) - (5a), the following reduced-form model can be estimated by inserting (5a) into (4a):[13]

(6a) $$Y_i^k = C_i^k \underbrace{(\alpha + \beta H)}_{\psi} + \underbrace{v\beta + u_i^k}_{\varepsilon_i^k}$$

In order to measure ex-ante IO, we are only interested in estimating the total effect $C_i^k \psi$ of circumstances without distinguishing their direct effect $C_i^k \alpha$ from their indirect effect $C_i^k \beta H$

[12] Circumstances are economically exogenous but econometrically endogenous because they are correlated with unobserved efforts encompassed in ε_i^k. Additionally, notice that Bourguignon et al. (2007a) and Ferreira and Gignoux do not apply subscript or superscript on most of the variables considered here, including the error term *v*. Applying a proper subscript or superscript on this term will be a topic of discussion in section II.3.

[13] Notice that in the original model from Bourguignon et al. (2007a), the log of income is considered instead of actual income (and consequently, the counterfactual income distributions from Ferreira and Gignoux (2011) described below are in linear-exponential form).

through efforts. This solves the previous problem of endogeneity arising from omitted effort variables in the structural model (4a)-(5a). Hence, no effort variable needs to be observed for the purpose of consistently estimating $C_i^k\psi$ in the reduced-form model (6a).

Ferreira and Gignoux (2011) have shown that the parametric approach from Bourguignon et a. (2007a) is close to the non-parametric approach from Checchi and Peragine (2010) in the sense that $\{\hat{\mu}_i^k = C_i^k\hat{\psi}\}$ and $\{\hat{v}_i^k = \overline{C}\hat{\psi} + \hat{\varepsilon}_i^k\}$ can be interpreted respectively as the ex-ante smoothed and standardized income distributions parametrically estimated.[14] Indeed, income differences in the counterfactual distribution $\{\hat{\mu}_i^k\}$ are due to different circumstance across individual (since C_i^k is the sole source of variation in this counterfactual), while income differences in the counterfactual $\{\hat{v}_i^k\}$ are not due to different circumstances across individuals. Notice that if equation (6a) does not suffer from omitted effort variables bias, it does however suffer from omitted circumstance variables bias, such that $\hat{\psi}$ is a biased estimator of ψ . Fortunately, Ferreira and Gignoux (2011) have proven that this omitted circumstance variables bias does not imply that the measured IO merely corresponds to the IO corresponding to observed circumstances only: it can be interpreted more generally as a lower-bound estimate of the true IO resulting from all possible circumstances, either observed or unobserved. The reason is that introducing new circumstances variables into the vector C_i^k will increase the variation of income accounted by $\{\hat{\mu}_i^k = C_i^k\hat{\psi}\}$. This result holds true when estimating IO non-parametrically as well. Intuitively, introducing new circumstance variables (or new categories within each circumstance) is equivalent to partitioning the population into more types. From equation (1), simple computation shows that as the number of types *K* increases, the between type component of the Theil L index increases too.[15]

Ferreira and Gignoux (2011) argue that applying the Theil L index over the smoothed income distributions $\{\hat{\mu}_i^k\}$ or $\{\mu_i^k\}$ are merely alternative procedures for measuring the same thing so long as the obtained measure of IO is interpreted as a lower-bound estimate of the true IO: the former is parametric estimation of the latter.[16] They further argue that "although

[14] A hat denotes an estimator of the corresponding true parameter or counterfactual distribution. $\overline{C}$ denotes the vector of average circumstances across all individuals.

[15] The between type component of the Theil L index will converge towards overall income inequality as the size of each type reduces up to one and the number of types increases up to N, which would amount to confound types with individuals in fact.

[16] And likewise for the standardized income distributions $\{\hat{v}_i^k\}$ and $\{v_i^k\}$, so long as the measured IE is interpreted as an upper-bound of the true IE.

[the parametric and nonparametric measures of IO] are estimates for the same path-independent measures, the fact that they are estimated parametrically, involving linear functional form assumptions, means they are not exactly identical" (p.635). The two methods indeed yield close results, which are nevertheless statistically different from each other at the 1% level, as shown by table 1. In the next subsection, we are going to consider the reason why the two methods are in fact not exactly equivalent and we are going to try to reconcile the two approaches.

Table 1: Checchi and Peragine's (2010) versus Ferreira and Gignoux's (2011) measures of IO

Approach	Mean (IO4)	Std. Err. (IO4)	F test
Checchi and Peragine	.0245	.00123	-
Ferreira and Gignoux	.0258	.00141	47.28***
Ferreira and Gignoux (ln)	.3218	.00476	4886.19***

Note: IO4 corresponds to IO based on non-parental circumstances solely, with four types considered (see the empirical part of the paper). Mean(IO4) corresponds to the average level of IO computed across 80 year-states observations. "Ferreira and Gignoux (ln)" corresponds to the case where equations (3), (4a) and (6a) are in log-linear form and where, consequently, the parametric smoothed income distribution is computed as $\left\{\hat{\mu}_i^k = \exp(C_i^k\hat{\psi})\right\}$. *The null hypothesis of the F test is that the mean IO computed according to Ferreira and Gignoux's method is equal to the mean IO computed according to Checchi and Peragine's method. *** indicates that the difference between the two means is significant at the 1% level.*[17]

3. *A synthetic approach*

This subsection shows that by reconsidering C_i^k in equation (6a) as a unique categorical variable denoting the type k individual i belongs to – whose coefficient will represent the simultaneous effect of all observed circumstance variables – rather than as a vector of circumstance variables, then the parametric method from Ferreira and Gignoux (2011) yields a measure of IO that is exactly equivalent to the one from Checchi and Peragine (2010). Indeed, as Ferreira and Gignoux (2011) explain it, the fact that their measure of IO is not exactly equivalent to the one from Checchi and Peragine (2010) is due to the linear form assumption of equation (6a). However, the important point to notice here is that the difference between the two methods does not lie in the linearity in parameters of equation (6a). Rather, it

[17] Notice that Ferreira and Gignoux's method is even more distant from Checchi and Peragine's method in the case where the log of income – instead of actual income – is considered in equations (3), (4a) and (6a). This suggests that it is not appropriate to consider individual income as a Cobb-Douglas function of efforts and circumstances.

lies in the linearity in predictors because equation (6a), as it stands, estimates the effect of each circumstance on income while holding other circumstances variables fixed. In contrast, the fact for an individual of belonging to a given type k is a function of all circumstance variables considered simultaneously.

First, notice that circumstances are categorical variables (I will discuss later that they are categorical *by nature*). Hence, C_i^k in Ferreira and Gignoux (2011) is a vector of categorical circumstances variables C_i^j (eg. gender or skin color). Let's denote by X^j the number of categories considered within C_i^j and by x_i^j a particular category within C_i^j (eg. male for the gender circumstance variable). Using standard dummy coding for categorical variables, each circumstance variable C_i^j can thus be split into *(X^j-1)* dummy variables x_i^j. Imagine further that only one circumstance variable C_i^j is observed. Making explicit the intercept term δ in equation (6a) yields:

(6b) $$Y_i^j = \delta + C_i^j \psi^j + \varepsilon_i^j = \delta + \sum_{x^j=1}^{X^j-1} \psi^{x^j} x_i^j + \varepsilon_i^j$$

where δ is a scalar representing the mean income of individuals belonging to the omitted category, ψ^{x^j} is a scalar representing the difference between the mean income of individuals belonging to category x_i^j and the mean income of the omitted category, and ψ^j is a vector whose elements are the coefficients ψ^{x^j}. In this framework, since we have only one circumstance variable, belonging to category j means belonging to type j, and the smoothed income distribution parametrically estimated $\left\{\hat{\mu}_i^j = \hat{\delta} + C_i^j \hat{\psi}^j\right\}$ from Ferreira and Gignoux (2011) exactly corresponds to the one from Checchi and Peragine (2010).

Now, imagine once again that several (J) circumstance variables are observed. Let's rewrite equation (6b) in order to show explicitly the effect of each circumstance variable:

(6c) $$Y_i^k = \delta + C_i^k \psi + \varepsilon_i^k = \delta + \sum_{j=1}^{J} C_i^j \psi^j + \varepsilon_i^k = \delta + \sum_{j=1}^{J} \sum_{x^j=1}^{X^j-1} \psi^{x^j} x_i^j + \varepsilon_i^k$$

Because equation (6c) is a multiple linear equation (contrarily to equation (6b) which is a simple linear equation), each coefficient ψ^{x^j} needs now to be interpreted *ceteris paribus*, that is, holding all other circumstance variables different from j fixed. Consequently, ψ^{x^j} cannot anymore be interpreted as the difference between the mean income of individuals belonging to category x_i^j and the mean income of the omitted category (whether this category is defined as the omitted category in variable j or in all circumstance variables). The smoothed income distribution from Ferreira and Gignoux $\left\{\hat{\mu}_i^k = \hat{\delta} + C_i^k \hat{\psi}\right\}$ is not anymore exactly equal to the one from Checchi and Peragine $\left\{\mu_i^k\right\}$ because it represents the effect of one circumstance variable at a time, while the others are held fixed.[18] To illustrate this, imagine that we seek to measure the extent of IO resulting from two circumstance variables, namely region of birth (South/North) and skin color (Black/White), in the US at the time of the Civil War.[19] It makes intuitive sense that the fact of being Black was much more disadvantageous for those people born in Southern States that were practicing slavery, independently from the fact that Northern States were on average wealthier than Southern States. The smoothed distribution $\left\{\hat{\mu}_i^k\right\}$ would miss this point: it would overestimate the value of the opportunity set of a Black person born in the South and underestimate the one of a Black person born in the South. Likewise, $\left\{\hat{\mu}_i^k\right\}$ would reflect the fact that being born in the South is on average disadvantageous but it would miss the point that it might be an advantage for White people.

What we want to estimate is the mean income of each type. Since a type gathers people that belong to the same category for each circumstance variable, what we are looking for is the interaction effect of all circumstance variables, not the separate effect of each circumstance variables while holding others fixed. To do so, we need to define a single categorical variable c_i^k that indicates the type k individual i belongs to (the way each type is coded has no importance). Substituting upper case C_i^k by lower-case c_i^k in the model of equations (4a) to (6a) and making explicit the intercept δ yields:[20]

$$Y_i^k = \delta + c_i^k \alpha + E_i^k \beta + u_i^k \tag{4d}$$

$$E_i^k = Hc_i^k + v \tag{5d}$$

$$Y_i^k = \delta + c_i^k \underbrace{(\alpha + \beta H)}_{\psi} + \underbrace{v\beta + u_i^k}_{\varepsilon_i^k} \tag{6d}$$

where c_i^k is a vector of *(K-1)* dummy variables x_i^j and δ is the mean income of the omitted type. All the desirable properties of the model (4a) - (6a) are shared by (4d) - (6d) in the sense that the latter: i) allows to represent individual income as function of efforts and

[18] This explains why the sum of IO and IE estimated following the method of Ferreira and Gignoux (2011) is not exactly equal to total inequality even when using the Theil L index, which is path-independent.

[19] This example is not taken for controversial ends. It is only taken because it is illustrative. Additionally, it does not seek to be historically exact.

[20] In technical terms, the reduced-form model (6d) corresponds to a non-additive multivariate ANOVA model.

circumstances; ii) distinguishes conceptually the direct and indirect effects of circumstances; iii) gives the full effect of circumstances ψ which does not suffer from omitted effort variables; iv) yields a measure of ex-ante IO that can be interpreted as a lower-bound. In addition, it has the additional property of allowing to estimate parametrically a smoothed income distribution $\left\{\mu_i^k = \hat{\delta} + c_i^k \hat{\psi}\right\}$ that is now exactly equal to the one from Checchi and Peragine (2010) instead being an approximation of it. The formula for this new smoothed income distribution parametrically estimated corresponds to the one from Ferreira and Gignoux (2011) in the sense that both are the predicted values of income regressed on circumstances. The only difference is that the circumstance variables considered is not same: a vector of circumstance variables in Ferreira and Gignoux (2011) against a unique variable representing the interaction effect of all circumstances. Regarding the standardized income distribution, an exact equivalent to Checchi and Peragine (2010) cannot be obtained fully parametrically. It can only be obtained semi-parametrically as $\left\{v_i^k = Y_i^k \dfrac{\mu}{\hat{\delta} + c_i^k \hat{\psi}}\right\}$.

The parametric method from Ferreira and Gignoux (2011) is said to have the advantage of being more robust to small sample sizes compared to the non-parametric one from Checchi and Peragine (2010). From a statistical standpoint, this is logical since the former consists in splitting the population into the categories defined by one circumstance variable only (while the others are held fixed) and repeating the operation for each circumstance variable. The number of observations per category when one circumstance variable at a time is considered is necessarily larger than the number of observations per type (where all categories from all circumstances are considered simultaneously). As a matter of fact, the number of observations per type decreases geometrically as the number of circumstances (or the number of categories per circumstance) increases, which could prevent a consistent estimation of the mean income of types with very few observations. However, I do not believe the problem of having very few observations per type to be mainly a statistical one. Rather, the problem has to do with the categorical *nature* of the circumstance variables. To the exception of a very few papers,[21] circumstances have always been treated as

[21] O'Neill et al. (2000), Nillson (2005) and Figueiredo et al. (2010) treat parental income as a continuous circumstance variable. They seek to estimate opportunity sets defined as the distribution of children's income conditional on the distribution of parents' income (the approach does not hinge on considering parental income as the sole circumstance variable). For instance in O'Neill et al. (2000), the fact that children in the US whose parents are poor (ie. the parents are at the 25th percentile of their income distribution) need to be at the 70th percentile of their income distribution in order to reach the average income of all children in the population indicates the existence of IO. This does not invalidate my point about the categorical nature of the circumstance

categorical variables but only for a matter of empirical convenience, not because they are categorical *by nature*. This is quite well illustrated by Björklund et al. (2011, p.18) who categorize some continuous candidates for circumstances (such as parental income) but are concerned with the fact that this "ignores some within-type variation in circumstances and thus underestimates the importance of circumstances and overestimates the role of effort". I posit that whenever they are used as a circumstance, these non-discrete candidates should be treated as categorical variables without sharing this concern.[22] Indeed, treating circumstances as continuous variables amounts to increase the number of potential types considered up to infinity and the number of types actually observed up to *N*. Doing so leads the between-type component of inequality in equation (1) to increase up to total inequality (and the within-type component to decrease to zero), such that the obtained measure would not anymore be an IO one.

The choice of the categories in which circumstances are categorized is necessarily subject to arbitrary decisions, just in the same way as the choice of the circumstances variables themselves is arbitrary: are they really outside the sphere of the individual's control? do they have a significant impact on individual income? The categorization procedure should reflect what can reasonably be thought to affect the opportunity set faced by individuals. Imagine we want to include monthly parental income as a circumstance variable. Categorizing it in hundreds or thousands of dollars makes sense, categorizing it in dozens of dollars not so much: can an individual be said to be unfairly disadvantaged compared to another one just because his parents are earning a few dollars less per month? I do not believe so.[23] The problem with the categorization procedure is that increasing the number of categories must at some point become irrelevant, thereby leading to overestimate IO. Notice that this does not invalidate the lower-bound result from Ferreira and Gignoux (2011) so long as the categorization procedure is done cautiously.[24] To sum up: having very few observations per

variables because their approach is fundamentally different from the one considered in this paper. Their approach serves to test the null hypothesis of EO or to rank income distributions based on a stochastic dominance criterion in order to compare pairwise the degree of IO between two income distributions. It does not serve to measure numerically the degree of IO based on an inequality index. Put differently, their approach serves to estimate opportunity sets, not to value them.

[22] Most circumstances variables – such as race, birth place or gender – are in any case qualitative and could not be treated continuously even if we wanted to.

[23] In practice, one would seek to categorize parental income into quantiles rather than into income ranges. The reasoning remains however the same in the sense that defining very small quantiles becomes at some point irrelevant.

[24] The categorization procedure seems to be more easily subject to questionable decisions from the social scientist than the choice of the circumstance variables themselves. This pinpoints to the necessity of either: (a) assuming the existence of a political agreement on the appropriate list of circumstances and categories; (b) effectively reaching this agreement within the academia.

type should primarily be a concern because this indicates that the way circumstance variables are categorized is not relevant, and incidentally because this might prevent a consistent estimation of IO from a statistical standpoint.

Before turning to derive the conditions under which IO might be beneficial or detrimental to growth, let me tackle a final point that has been overlooked so far: the error term v in equations (5a) and (5d).[25] This term is described by Bourguignon et al. (2007a) and Ferreira and Gignoux (2011) as a well-behaved disturbance term accounting for unobserved effort and circumstance variables as well as other unobserved stochastic income determinants such as luck. The present paper posits that it can alternatively be seen as a measure of relative efforts. Indeed, E_i^k is necessarily a measure of absolute efforts since it is partly explained by circumstances. On the contrary, v accounts for the share of absolute efforts that is not explained by circumstances: it is a measure of relative efforts and can be seen as an idiosyncratic propensity to provide efforts. Hence v should be written as v_i since it is proper to individual i.[26] This measure of relative of efforts is necessarily well-behaved (ie. $E(v_i \mid c_i^k) = E(v_i) = 0$) since it is assumed in the theory of IO that one's individual propensity to provide efforts (that is, one's merit) is unconditional on one's type. It is negative for those individuals whose absolute level of efforts is lower than the mean level of absolute efforts provided by the other members of their corresponding type, it is positive otherwise.

Notice that the parametric estimate $\left\{\mu_i^k = \hat{\delta} + c_i^k \hat{\psi}\right\}$ corresponds to the ex-ante smoothed distribution because $E(v_i \mid c_i^k) = E(v_i) = 0$. Indeed, this estimate corresponds to the expected value of income conditional on the type k individual i belongs to in equation (6d). Further, notice that a parametric definition of ex-ante EO is given by:

(7) $$\psi = 0,$$

meaning that the total effect of circumstances on individual income is null and that, consequently, the mean income of all types is the same: $\mu_i^k = \delta, \forall i,k$. Assuming that the relationship between absolute, relative efforts and circumstances is linear in predictors (ie. there is no interaction term $c_i^k \cdot v_i$ in equation (5d)), a parametric definition of ex-post EO is

[25] The forthcoming discussion holds true for both models (4a)-(6a) and (4d)-(6d).

[26] To be fully exact, v should be written as $v_i^k = v_i + \vartheta_i^k$, where v_i is the measure of relative effort while ϑ_i^k accounts for all other unobserved determinants of absolute efforts that may either be proper to individual i, either be proper to his type k.

given by:[27]

(8) $$\alpha = 0 \text{ and } H = 0,$$

meaning that circumstances neither have a direct effect on income, nor an indirect one through efforts (recall that $\psi = \alpha + \beta$). Clearly, we see, as proven by Fleurbaey and Peragine (2013), that the conditions for achieving ex-post EO are more demanding than those for achieving ex-ante EO: the former requires circumstances not to have whatsoever effect on income while the latter allows circumstances to have both a direct and an indirect effect so long as the two effects offset each other. Since ex-post EO implies ex-ante EO, it follows that ex-ante IO implies ex-post IO ("A implies B" => "non-B implies non-A"). In other words, a measure of ex-ante IO is necessarily smaller than a measure of ex-post IO based on the same set of circumstances. Likewise, the variation of IO in a growth regression will tend to be smaller when measured following the ex-ante approach. This paper consequently adopts empirically ex-ante measures of IO in section IV because it will prove harder to find a significant impact of IO on growth – the hypothesis that this paper has originally been seeking to confirm but that is invalidated by the data, as we will see in section IV – using this measure whose variation tends to be smaller compared to an ex-post measure.[28]

4. *The theoretical impact of IO on growth*

Considering the impact of IO on growth is not an easy task. The question is basically the following one: does the way the pie is divided impact the size of the pie itself? This subsection studies the conditions under which IO *might be* beneficial or detrimental for aggregate income. It then presents some illustrative facts about Brazil in order to detect which conditions this country is fulfilling. Two points are worth noticing here. First, the analysis remains a static one in the sense that it studies the impact of an instantaneous change in IO on aggregate income. The reasoning could however be extended to the case where the impact of a gradual change in IO on growth is under scrutiny. Second, the facts about Brazil describing whether this country fulfills conditions under which IO might be beneficial or detrimental to growth are not intended to predict the impact of IO on growth in Brazilian states. Rather, they

[27] It is likely that this assumption does not hold in practice because individuals from advantaged types have a larger set of opportunities in the sense that they can obtain more easily the income they desire given their circumstances and idiosyncratic preference for efforts. Put differently, advantaged individuals with a low propensity to provide efforts have the opportunity to provide a very little amount of absolute efforts, while advantaged individuals with a high propensity to provide effort have the opportunity to provide a very large amount of absolute efforts. This is exactly what the interaction term $c_i^k . v_i$ would capture, were it inserted into equation (5d).

[28] Notice that Marrero and Rodríguez (2012a) find a significant effect of IO on growth using an ex-ante measure.

are intended to facilitate the interpretation of the econometric results from section IV.

The model of equations (4d)-(6d), as it stands, does not allow to study the impact of IO on aggregate income because the categorical variable c_i^k does not rank types according their mean income in this model: the way each type is coded – whether one type is assigned a higher or lower numerical value than an other one and the choice of the omitted category –has no importance at all for the purpose of estimating the smoothed income distribution in equation (6d). Now, in order to study the impact of IO on aggregate income, it is necessary to code each type according to its mean income and to treat c_i^k as a continuous variable in the following model:

(4e) $$Y_i^k = \delta + c_i^k \alpha + E_i^k \beta + u_i^k$$

(5e) $$E_i^k = Hc_i^k + F(c_i^k)^2 + v_i$$

(6e) $$Y_i^k = \delta + c_i^k \underbrace{(\alpha + \beta H)}_{\psi} + \beta F(c_i^k)^2 + \underbrace{v_i \beta + u_i^k}_{\varepsilon_i^k}$$

where absolute efforts E_i^k are now treated as a scalar rather than as a vector of effort variables for expositional simplicity purposes. Treating $c_i^k = \{1,2,...,K\}$ as a continuous variable growing integer by integer according to the rank of each type's mean income allows to insert the squared term $(c_i^k)^2$ into the model.[29] Because c_i^k is now *treated as* a continuous variable, it does not represent anymore the interaction effect of all circumstance variables as in model (4d)-(6d): it represents the rank of individual *i*'s opportunity set value (compared to the value of the opportunity set of individuals belonging to types other than *k*). c_i^k can seen as an input in the individual income function that represents the underserved advantage position of individual *i* compared to individuals from other types. Hence, treating c_i^k as a continuous variable in the model (4e)-(6e), which is not intended to estimate the smoothed income distribution, is not denying the categorical nature of the circumstance variables. Inserting the quadratic term $(c_i^k)^2$ serves to study the second order derivative of individuals' income with respect to the relative value of their opportunity set.

Now, defining aggregate income as the sum of individual incomes across the population ($\sum_{i=1}^{N} Y_i^k$) and considering two types – let's call them *a* and *b* with *a*>*b* meaning that

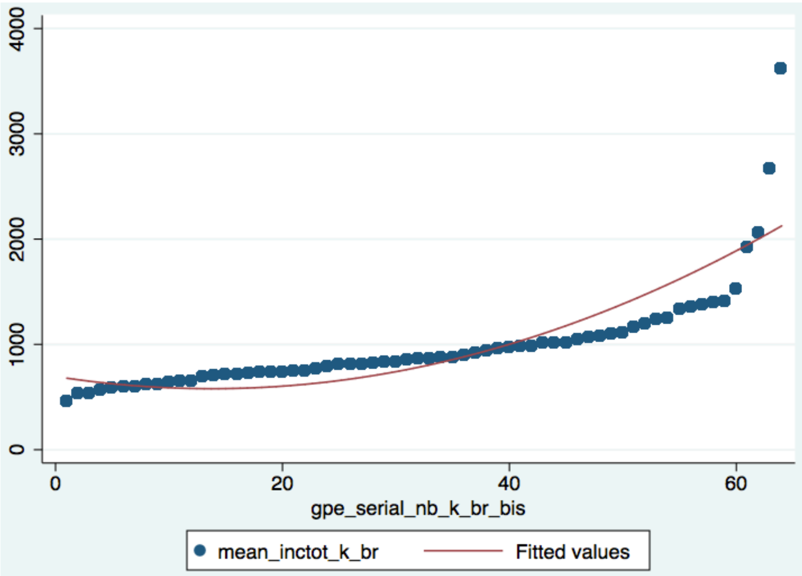
[29]Even though c_i^k is not continuous since it grows integer by integer, it is *treated as* a continuous variable. If c_i^k were treated as a categorical variable, then inserting $(c_i^k)^2$ would bring no additional information. In econometric terms, this would result in perfect muticollinearity between c_i^k and $(c_i^k)^2$.

the mean income of type a is greater than the mean income of type b – the following statement can be made: a reduction of IO obtained by improving marginally the situation of type b (ie. whose rank goes up of 1 unit) and by deteriorating marginally the situation of type a (ie. whose rank goes down of 1 unit) will lead to a reduction of aggregate income if F in equation (6e) is positive and to an increase of aggregate income if F is negative. Mathematically, the reason is that Y_i^k is monotonically increasing in E_i^k and c_i^k, and that it is convex in c_i^k when $F>0$ (and concave in c_i^k when $F<0$). Of course, this statement is not a proof of the effect of IO on aggregate income because it considers a special case of reduction of IO: for instance, it is possible to reduce IO by equalizing all types' mean income to the mean income of the overall population without affecting aggregate income, whatever the sign of F. However, this statement does pinpoint to conditions under which IO might be expected to have beneficial or detrimental effects on aggregate income: for instance if F is positive, it is impossible to equalize all types' mean income to the mean income of the median type – the one ranked $(K+1)/2$ when the number K of types is an odd number, which can be seen as the type that is not unfairly (dis)advantaged – without reducing the aggregate income.

Put differently, the quadratic term $(c_i^k)^2$ is inserted in equation (5e), and incidentally in equation (6e), in order to reflect the fact that the relationship between absolute efforts and circumstances might be nonlinear: individuals take their circumstances as given but, since they are aware of them, they chose their absolute effort level according to the marginal returns of their opportunity set's relative value (these marginal returns correspond to F). In consequence, when there are increasing marginal returns to opportunities ($F>0$), the existence of IO can be viewed as a misallocation of opportunities resulting in an aggregate underprovision of efforts. The analysis is in line with Marrero and Rodríguez (2012b) who develop a dynastic model where the accumulation of human capital – which I believe can be interpreted as a proxy for efforts – results from an individual decision to provide efforts given circumstances. In their model, disadvantaged individuals accumulate less human capital and since there are decreasing returns to human capital accumulation (not to be confounded with marginal returns to opportunities), higher degree of IO leads to lower than optimal aggregate level of human capital and hence lower economic performance. However, Marrero and Rodríguez (2012b) also show that when social mobility is included in the analysis, the relationship between IO and growth becomes in fact nonlinear and is actually positive in less developed countries.

Figure 1 shows the graphical illustration of equation (6e) in Brazil for the year 2000.[30] However, because there are so many observations in the sample, the mean income of each type is represented on the Y-axis instead of individual income.[31] The rank of each type (ie. c_i^k) is represented on the X-axis. We see that the form of the function is almost linear. Only a few observations concerning the most advantaged types drive the curve up and render it convex. I interpret this as an indication that the impact of IO on growth in Brazil is either null or slightly positive, which is confirmed by the empirical part of this paper.

Figure 1: Returns to opportunities **(**Brazil 2000)



III. Data and inequality figures

This section presents the empirical strategy designed to estimate inequality in Brazilian states. As previously explained, computed measures of IO and IE are based on the ex-ante approach. They follow the non-parametric framework from Checchi and Peragine (2010) and are computed accordingly to equation (1).

---

[30] The figure considers the whole Brazil, not Brazilian states. Notice that the same figures have been realized for the years 1991 and 1980. They are not shown because they look the same as the one displayed here.

[31] The reasoning developed in the preceding paragraphs can be extended similarly to the case where the mean income of each type is considered in equation (6e) instead individual income. I chose to represent individual income in equations (4e)-(6e) for purpose of consistency with the previous subsections.

1. Data on individual income and circumstances

In order to compute ex-ante IO indices, data on individual income and circumstances is required. IPUMS-International inventories and harmonizes census microdata from around the world. Regarding Brazil, it includes individual-level data on income and circumstances (skin color, gender, paternal occupation and paternal level of education) coming from the national demographic censuses ran by the *Instituto Brasileiro de Geografia e Estatística* (IBGE) in 1960, 1970, 1980, 1991 and 2000. The censuses are representative both at the national and state levels. Years 1960 and 1970 are excluded from the analysis because the former only reports income values expressed as broad income ranges and the latter does not contain information on skin color. Because each observation does not represent the same number of individuals, weights are taken into account thanks to the IPUMS variable *WTPER* in order to yield statistics that are representative of the population.[32]

One of the advantages of the Brazilian censuses is that their sample sizes are very large: they contain 5,870,467 observations in 1980; 8,522,740 in 1991; and 8,446,353 in 2000.[33] In order to focus on the active population, the analysis is restricted to people with positive reported income aged between 20 and 49, as is common in the literature. Because data on parental occupation and parental education is only available for those individuals living in the same household as their parents (who are not representative of the population as will be discussed later), two alternative samples need to be considered. The first one (sample A) gathers individuals with all non-parental circumstances (gender and skin color) observed and does not suffer from sample bias.[34] It serves to compute total inequality as well as IO and IE based on these sole non-parental circumstances. However, it would not be satisfactory to ignore altogether parental occupation and parental education from the analysis because these two circumstance variables have been assessed as the two most important ones explaining IO in Brazil (Bourguignon et al., 2007a; Ferreira and Gignoux, 2011). The problem here is that restricting the analysis to individuals whose parental circumstances are observed (in addition to the non-parental ones) would yield some biased measures of inequality because this sample

[32] In census data, the sample design includes clustering and stratification in addition to weights. Contrarily to weights, clustering and stratification only affects the estimated standard errors, not the point estimates. For the purpose of computing inequality non-parametrically, I was only interested in obtaining unbiased point estimates. Hence, clustering and stratification did not need to be taken into account. For more details about survey estimation and the IPUMS databases, see Davern and Strief (2008) and Cleveland et al. (2011).

[33] Because such large datasets are computationally very expensive, the three censuses were treated separately.

[34] Were also excluded from sample A those very few individuals (their number is negligible) whose non-parental circumstances are observed but whose parental circumstances are not observed for a reason other than selection into parental household (their parental circumstances are not observed because the desired census questions were not answered by the respondent). It would be more exact to refer to sample A as the sample gathering those individuals with no unobserved circumstances due to selection into paternal household.

(called sample A1) of individuals is not representative of the population due to selection into parental household. In order to treat this problem, a procedure of multiple imputations is adopted, as discussed in section III.2. Lastly, notice that it is common in the IO literature to restrict the analysis to household heads. I do not do so because this would have prevented me from including gender as a circumstance variable. Indeed, this would have resulted in another source of sample bias since household headship is endogenous to gender.

Table 2: Sample sizes by census year

	1980	1991	2000
Original IPUMS sample.	5,870,467	8,522,740	8,446,353
(fraction of overall Brazilian population)	(5%)	(5.8%)	(5%)
Aged 20-49 and with positive income reported.	1,443,967	2,260,049	2,415,269
Sample A: all non-parental circumstances observed.	1,439,672	2,255,578	2,401,739
Sample A1: all parental circumstances additionally observed.	228,815	360,358	375,629

Note: The table refers to the number of observations on individuals with observed characteristics, not to the number of individuals itself (since each individual is assigned a weight in census data). From the top to the bottom, each additional row imposes a further restriction on the sample considered.

Individual income corresponds to monthly total personal income from all sources.[35] It was preferred over labor earnings (also available through the IPUMS database) because total income depicts the full set of opportunities an individual is enjoying from while labor earnings are restricted to those opportunities related to the labor market. In the empirical IO literature, it is common to remove the effect of age on income. The reason is that age is neither seen as a circumstance nor as an effort: hence, we do not want it to explain income differences across individuals. In order to remove the effect of age, most papers restrict their analysis to individuals belonging to a narrow age group (eg. 30 to 40 years old). This has the disadvantage of reducing considerably the sample size (in addition to yielding statistics that are less representative of the overall population). Even though the IPUMS samples are very large, the number of observations on individuals living in the same household as their parents (ie. whose parental circumstances are observed) is less than one fifth the total number of available observations is sample A, as shown in table 1. Consequently, I do not further restrict the sample to a narrow age group because this might put at risk the multiple imputation

[35] Monthly income corresponds to the previous month's income for employees, or the average monthly income of the year for persons who did not work the previous month or otherwise had irregular or variable income.

procedure considered in section III.2.[36] Rather, I adopt a strategy inspired from Checchi and Peragine (2010) and Marrero and Rodríguez (2012a). First, income is regressed on age and age squared. Second, the predicted residuals of this regression are taken. Third, because residuals are centered around zero, they are added a constant term in order to match the minimum of the actual series. The obtained adjusted income represents the income earned by an individual of average age. Notice that Checchi and Peragine (2010) and Marrero and Rodríguez (2012a) regress income on (potential) experience instead of age in the first step. I do not follow them because experience is not a factor that should be removed. Experience (ie. the number of years of presence in the labor market) is the result of individual decisions (including length of study, for instance) and should thus be considered as an effort variable. Consequently, removing its effect would inappropriately reduce IE as well as IO (due to the indirect effect of circumstances through effort, as explained in section II).

Turning to the circumstance variables, the categorization procedure is done so as to reflect what can reasonably be thought to affect individuals' opportunity set, as explained in section II.4. Regarding the non-parental circumstances, skin color is coded into two categories: White and Asians people; others (Black, Mixed-race and Indigenous people). The gender variable is naturally coded into two groups: male; female. Regarding the parental circumstances, only father's information is considered because using mother's information as well would have complicated too much the multiple imputation procedure. Father's information is obtained thanks to a constructed family interrelationship variable (*POPLOC*) developed by the IPUMS-International. It indicates the location within each household of every person's father. The design of this variable is based on a complex core algorithm and accounts for the peculiarities of each Brazilian census. This algorithm is highly reliable: the IPUMS pointers agree with direct reports of family interrelationships more than 98% of the time (Sobek and Kennedy, 2009). Notice that this variable identifies social relationships (such as stepfather and adopted father) as well as biological relationships. This is highly relevant for the purpose of measuring IO: a person is relatively (dis)advantaged based on his inherited social characteristics, not on his "genetic" characteristics (apart from ethnicity, to the extent that there exist some racial discrimination, but this is accounted by the skin color variable).

[36] The problem is that this might further reduce the proportion of missing observations on parental circumstances. There would be no problem with respect to the multiple imputation procedure if I were to restrict the analysis to people aged between 20 and 30 years old because individuals living in the same household as their parents are overrepresented in this age group. However, doing so would not be appropriate because I want to focus on the active population. Lastly, notice that total income is top-coded at 9,999,997 Cr$ for the year 1991. This is however not a source of concern because the proportion of individuals above this threshold is negligible (95 observations in sample A, which is less than 0.005% of individuals (weights taken into account) of sample A).

Notice additionally that information based on questions directly asked to the father is more precise than information based on questions asked to the child. Two parental circumstances are considered: father's educational attainment and occupational status. Father's education is coded into four categories: no education, pre-school or literacy courses; primary 1-4; primary 5-8; secondary or more.[37] Father's occupation is based on the original IPUMS variable *OCCISCO*, whose universe concerns people in the labor force, and which is coded according to the ten major categories of the 1988 *International Standard Classification of Occupations* (ISCO) scheme.[38] *OCCISCO* is recoded into four categories whose choice is based on two criteria: i) not having one category disproportionately larger or smaller than the other ones; ii) having some consistency within each category (in the sense that each category gathers individuals with comparable occupational status). The first category includes fathers that are not in the universe of the *OCCISCO* variable. Most of them are in fact retired.[39] The second category includes: skilled agricultural and fishery workers; elementary workers. The third category includes mainly non-agricultural manual workers: plant and machine operators and assemblers; crafts and related trades workers; service workers and shop and market sales. The fourth category includes mainly desk workers and skilled professions: clerks; technicians and associate professionals; professionals; legislators, senior officials and managers.[40]

2. *Sample bias and multiple imputation*

Though we can be confident that the IPUMS variable *POPLOC* matches appropriately an individual to his father, obtaining father's information thanks to this pointer variable poses some serious empirical problems. Indeed, paternal circumstances are only observed for those

---

[37] In 1971 Brazil converted from a 4-4-3 educational system (4 years of primary school, 4 years of middle school or junior high school, and 4 years of secondary school) to an 8-3 system (8 primary, 3 secondary). Hence, the last for years of primary school (primary 5-8) correspond to middle school under the old system.

[38] *OCCISCO* indicates the occupation of the father at the time of the census. One might argue that children's achievement impacts the occupational status of the father. In this case, *OCCISCO* would not be economically exogenous to children's income and would thus not be a true circumstance variable. Inequality measures based on non-parental circumstances serve as a robustness check against this possibility. Notice that I could have considered father's educational attainment as the sole parental circumstance. I chose to include father's occupation as a circumstance variable nevertheless because: i) I believe the causality to go mainly in the opposite direction (from the father to the children); ii) as already explained, father's occupation has been assessed to be an important factor explaining IO in Brazil. Consequently, the benefit of including father's occupation as a circumstance outweighs the cost of not including it.

[39] Another disadvantage of having *OCCISCO* indicating the occupation of the father at the time of the census is that retired fathers cannot be classified in a more precise occupational category.

[40] Not classified responses were grouped into the lowest skill category (ie. the second category), as is common in the literature. The last ISCO group not yet described is "armed forces". It does not seem to match any of the four categories described above. I make the arbitrary decision to include it into the fourth category because it is the smallest one in terms of sample size. In any case, this group of people is negligible: less that 0.06% of individuals of sample A (weights taken into account).

individuals living in the same household as their father. Because these individuals are not representative of the population, computing IO on this subsample (sample A1) may yield inequality measures suffering from sample bias as already mentioned. Indeed, an income distribution provides two crucial pieces of information for the purpose of measuring IO: the income and population shares of each type, as made evident by equation (1). The bias does not result solely from the fact that individuals with observed parental circumstances are not representative of the population in terms of income and circumstances: if all types were affected in the same proportion, then computing IO on this subsample would not be a problem.[41] In order to see whether this is the case, inequality measures based on non-parental circumstances (which are observed for everybody) can be compared over samples A and A1. Table 3a shows that these three figures computed over sample A1 are significantly different at the 1% level from those computed over sample A, which are not subject to sample bias. This means that IO measures based on non-parental circumstances and computed over sample A1 are suffering from sample bias due to selection into paternal household. Consequently, IO measures based on both parental and non-parental circumstance computed over sample A1 will, *a fortiori*, suffer from sample bias as well.

Table 3a: Evidence of sample bias (inequality measures computed over sample A versus A1)

	IO4	IE4	Total inequality
Sample A	.0245 (.0012)	.2973 (.0042)	.3218 (.0048)
Sample A1	.0062 (.0007)	.1287 (.0030)	.1348 (.0034)
F test	190.43***	2647.50***	2175.58***

*Note: IO4 and IE4 correspond respectively to IO and IE based on non-parental circumstances solely (where four types are considered). The table shows the mean of these inequality measures across 80 year-states observations. Standard errors are in parentheses. The null hypothesis of the F test is that the mean across year-states of an inequality measure computed over sample A is equal to the mean across year-states of the corresponding inequality measure computed over sample A1. *** indicates that the difference between the two means is significant at the 1% level.*

This paper adopts a procedure called "multiple imputation" (MI) in order to treat this problem of sample bias. It is a modern simulation-based statistical technique developed by Rubin (1977) that is especially designed for handling missing data. Newman (2003, p. 334) defines it as "a procedure by which missing data are imputed several times (e.g. using regression imputation) to produce several different complete-data estimates of the parameters.

[41] In technical terms, sample A would be considered as a replication of sample A1.

The parameter estimates from each imputation are then combined to give an overall estimate of the complete-data parameters as well as reasonable estimates of the standard errors." Two main types of variables are considered in MI: those with missing values that are to be treated are called "imputed variables" (father's education and occupation here); the other ones are called "regular variables".[42] In practice, MI consists of three steps. The first step, called imputation step, consists in generating several completed datasets. In these imputed (ie. completed) datasets, the missing observations on the imputed variables are replaced by their predicted values. These predicted values are obtained using all available information on both the imputed and regular variables. In order to account for the uncertainty caused by estimating missing data (ie. the imputed values are not the true values of the missing observations, they are a plausible versions of it), a random source of variation is introduced between each imputed sample.[43] The second step, called completed-data analysis step, consists in performing the desired estimation analysis separately on each imputed dataset. The last step, also called pooling step, consists in combining the results into a single set of unbiased parameter estimates.

The imputation step relies in practice on an imputation model where the dependent and independent variables are respectively the imputed and regular variables. Since the two imputed variables considered here – father's education and occupation – are qualitative, the chosen imputation model is a multinomial logit.[44] Notice that the choice of the predictors in this imputation step is much less demanding than in the first step of a Heckman procedure because the imputation model is not intended to model the reason why some observations are missing. Rather, the predictors are intended to be correlated with the missing data and causes of missingness in order to maintain the overall variability in the population while preserving the relationships between the variables of interest in the analysis phase (Wayman, 2003). Individual income and non-parental circumstances are included because they are explored in the analysis phase.[45] Age, a categorical variable for marital status, a dummy for rural or urban

[42] There exists also a third type of variables called "passive variables". They are not considered here because they are of no relevance in this paper. See Allison (2001), Wayman (2003) or Baraldi et al. (2010) for more details about the MI procedure.
[43] The random seed used in this paper is five. Several other seeds have been tried and the obtained inequality measures only change negligibly, which indicates that the imputation procedure has been implemented appropriately.
[44] Missing values on father's education and occupation concern exactly the same observations (ie. observations on individuals not living with their father). This joint pattern of missingness is called "monotone". This simplifies the process of imputation from a statistical standpoint: the multivariate imputation task can be formulated as a sequence of independent univariate (conditional) imputation tasks.
[45] Age is one of the most important variable explaining selection into paternal household. Consequently, the income considered in the imputation model is not adjusted for the effect of age. Income is then adjusted between

status, and a dummy for native- or foreign-born are included because they are likely to be correlated with the fact of living or not with one's father. Lastly, the variables describing the sample design of the censuses (the weight variable *WTPER*, and the primary sample unit *SERIAL*) are included, as strongly recommended in the literature. The validity of MI relies in theory on an infinite number of imputations. The procedure is however known to have good statistical properties with a finite number of imputations. Rubin (1987) derives a formula measuring the "relative efficiency" of a MI procedure with a finite number of imputations compared to the one with an infinite number of imputations.[46] Based on this measure, I choose to implement twenty imputations, yielding a relative efficiency of more than 99% for each the three years 1980, 1991 and 2000 considered.[47] Turning to the second step of the imputation procedure, measures of IO and IE based on both parental and non-parental circumstances have been computed on each imputed dataset separately following the non-parametric framework of Checchi and Peragine (2010). Lastly in the pooling step, I have followed Rubin's rule (1987) by computing the final inequality measures as the average of those estimated on each imputed dataset.

One of the main assumptions underlying the imputation procedure is that data be "missing at random" (MAR). This terminology is quite confusing because it does not describe a pattern of missingness that is random at all in fact. Allison (2001, p.4) defines the MAR assumption as a situation where "the probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis." Put more simply, it means that, given the observed data, the missingness mechanism does not depend on the unobserved data (otherwise, the imputed datasets would still, logically, suffer from sample bias). Unfortunately, this assumption is not directly testable in practice because the test would require the missing values to be observed. This strengthens the need to run a sensitivity analysis in order to test whether the whole imputation procedure (including the validity of the

---

the first and second steps of the imputation procedure following the method described in the first part of this section. Notice that the imputation model worked poorly when the income considered was adjusted for age, which makes sense and which I interpret as a good sign regarding the validity of the imputation model.

[46] The formula is: $\left(1+\frac{\gamma_0}{m}\right)^{-\frac{1}{2}}$, where γ_0 denotes the fraction of missing observations (ie. the ratio of the number of missing observations to the total number of observations) and *m* the number of imputed datasets.

[47] Given the very large size of the original IPUMS samples (see table 1), creating twenty imputed implies to manage several dozens of millions of observations for each year considered. The statistical software used in this paper (Stata/SE 11.2) cannot handle such a large number of observations. Consequently, the whole imputation procedure was done on a 10 percent random draw of the original IPUMS samples for the years 1991 and 2000. The smaller size of the 1980 IPUMS sample allowed to extract a 20 percent draw. The random draws were obtained thanks to the IPUMS variable *SUBSAMP*, which takes into account the design proper to each of the three censuses.

imputation model) has worked properly.[48] A methodology comparable to the one implemented to show that inequality measures computed over sample A1 were suffering from sample bias is adopted here. Inequality measures based on non-parental circumstances and computed over sample A, that are known to unbiased, are compared to the same ones computed over the imputed samples. Table 3b indicates that the results of the imputation procedure are satisfactory, though not perfect. The imputation procedure works very well regarding IO: the F test strongly fails to reject the null hypothesis that IO computed over the imputed samples is equal to IO computed over sample A (ie. the two measures are not statistically different from each other). Results about total inequality and IE computed over the imputed samples are less convincing: the F test indicates that they are not statically different form their corresponding measures computed over sample A but only at the 5% level. Nevertheless, this does not dismiss altogether the imputation procedure because the difference between the two types of measures only occurs at the third decimal place, which is not *economically* significant. Moreover, let me point out that the main variable of interest is IO, not IE or total inequality.[49] Lastly, in order to make sure that the empirical results from section IV are not driven by the fact that the MI procedure has not worked perfectly, all regressions in this section will be run alternatively on the inequality measures: i) based on both parental and non-parental circumstances, obtained thanks to the MI procedure (they will be referred to as *IO64* and *IE64* because they consider two categories per non-parental circumstance variable and four categories per circumstances variables, yielding a total of 64=2*2*4*4 types); ii) those based over non-parental circumstances only and computed over sample A (they will be referred to as *IO4* and *IE4* because they consider 4 types in total).

Before turning to assess empirically the impact of IO on growth, let me briefly describe the inequality indices computed in each year-Brazilian states and their evolution over time. Table 4 in the appendix shows the measures of total inequality, *IO4* and *IE4* computed over sample A as well as the measures of *IO64* and *IE64* computed over the imputed samples. Notice that the *IO64* measures are much higher than the *IO4* measures. This is normal because, as explained in section II, the more circumstance variables considered, the higher IO. The share of total inequality accounted by IO corresponds on average to 8% when considering

[48] Even though Little (1995), Rubin, (1996) and Graham et al. (1997) point out that the MI procedure is quite robust to departures from its underlying assumptions. Notice additionally that the MI procedure implemented here is more robust than the "textbook" MI procedure because I am only interested in estimating unbiased point estimates in order to measure inequality non-parametrically, not in drawing statistical inferences from the individual-level datasets.

[49] As a matter of fact, total inequality computed over the imputed samples will never be used in section 4 because the one computed over sample A corresponds t the "true" total inequality.

Table 3b: Imputation procedure (sensitivity analysis)

	IO4	IE4	Total inequality
Sample A	.0245 (.0012)	.2973 (.0042)	.3218 (.0048)
Imputed samples	.0248 (.0013)	.2920 (.0048)	.3168 (.0053)
F test	0.31 (p-val: 0.5804)	3.60* (p-val: 0.0615)	2.89* (p-val: 0.0933)

*Note: IO4 and IE4 are based on non-parental circumstances solely. The table shows the mean of the inequality measures across 80 year-states observations. Standard errors are in parentheses, unless otherwise stated. The null hypothesis of the F test is that the mean across year-states of an inequality measure computed over sample A is equal to the mean across year-states of the corresponding inequality measure computed over the imputed sample. * indicates that the difference between the two means is significant at the 10% level.*

IO4 and to 30% when considering *IO64*. This share of 30% is much more in line with previous studies about Brazil: for instance, Bourguignon et al. (2007a) find similar results of around one third in Brazil. All figures of inequality have decreased over the 1980-2000 period but the share of IO to total inequality has remained roughly constant (it has decreased during the 1980s and risen up to its previous level at the end of the 1990s). This means that all figures of inequality have decreased roughly in the same proportions. In comparison, Brazilian states have experienced an average growth rate of GDP per capita of 20%, 4% and 19% over the periods 1980-1989, 1991-2000, and 2000-2009 respectively. As an illustration, figure 2 in the appendix shows the graph of the simple linear regression of *IO64* on growth: the slope of the regression line is negative but insignificant.

IV. Empirical investigation of the growth-inequality relationship

This section investigates empirically the effect of IO on growth. As mentioned in the introduction, the only paper having studied this relationship so far is Marrero and Rodríguez (2012a), henceforth MR (2012a). They find that IO has a negative effect on the growth rate of US states, while has IE has a negative effect. They find that the effects of IO and IE are significant and robust to a battery of econometric specifications, contrarily to the one of total inequality which is found to be positive but whose significance is not robust. The aim of this section is to provide an independent replication of MR (2012a). To this end, I try to stick to their benchmark specifications as closely as possible given the availability of data: this section follows MR (2012a), unless otherwise stated.[50] In terms the results that I originally expected, the "inequality as cholesterol hypothesis" confirmed empirically by MR (2012a) seemed promising to me in order to explain the inconclusiveness of the growth-inequality literature. However, several clues from section II.4 indicate that the relationship between IO and growth might be positive, or at least not significantly different from zero, in the case of Brazil: i) the theoretical model from MR (2012b) suggests that the relationship between IO and growth might be positive in developing countries; ii) individual income displays increasing marginal returns to opportunities in Brazil (see figure 1).

1. Model specification

The empirical analysis is restricted to three decades, going from 1980 to 2009. It considers year-Brazilian states as the units of analysis, for a total of 80 observations since Brazil is made of 27 states including the Federal District (there are only 26 states in 1980 because the state of Tocantins was formed out of Goias in 1988). The analysis is highly comparable with Marrero and Rodríguez (2012a) in terms of sample size since they consider 26 US states across three decades going from 1970 to 2000. The following benchmark growth regression model is estimated with pooled-OLS, fixed and random effects:[51]

$$GY^{i}_{t,t+9} = \alpha + \beta' INEQ^{i}_{t,t+9} + \gamma Y^{i}_{t} + \lambda' X^{i}_{t} + \delta' R^{i} + \varphi' T_{t} + \varepsilon^{i}_{t} \tag{9}$$

where robust standard errors clustered by state are considered. $GY^{i}_{t,t+9}$ denotes the growth rate of GDP per capita of state *i* between time *t* (ie. 1980, 1991, or 2000) and time *t+9*. This growth rate is not multiplied by 100, as in MR (2012a). This does not impact the significance of the results but the coefficients on the explanatory variables should be multiplied by 100 in order to be compared with MR (2012a). A nine-year period is considered instead of a ten-year period as in MR (2012a) because data on GDP at the state level is not available in 2010. Lagged GDP per capita Y^{t}_{i} is included in order to control for conditional convergence across states. X^{i}_{t} is a set of control variables. They are measured at the beginning of the period in order to reduce their potential endogeneity. R^{i} and T_{t} control for regional and time fixed-

[50] Notice that the fact that my empirical results infirm those of MR (2012a) has strengthened the need to stick to their benchmark specifications.

[51] The random effects model is not reliable due to the unrealistic assumption that the time-invariant error term is uncorrelated with the independent variables. Results from this model are nevertheless shown for sake of comparison with MR (2012a). Notice additionally that MR (2012a) implement system GMM model. This model is not considered here because MR (2012a) do not provide information about the lags they use in this model (this information is necessary for the replication of this model). Additionally, although the pooled-OLS, FE and RE models are less efficient in taking care of the endogeneity of inequality to growth, they allow a straightforward interpretation of the results and an easy comparison with MR (2012a).

effects respectively. The time fixed-effects control for policies or economic shocks that are common to all Brazilian states, such as the period of high inflation in 1992.[52] Lastly, $INEQ_i^t$ denotes a set of inequality measures that will be described later. They are considered at the beginning of the period because I seek to study the medium-term effect of inequality on growth (in addition to reducing the endogeneity of inequality to growth). To the exception of fertility (as will be explained below), all the variables other than inequality in equation (9) come from Ipeadata, a website managed by the *Institute of Applied Economic Research* (Ipea) whose main purpose is to provide access to Brazilian economic data. All variables measured in monetary units are expressed in real terms (R$ de 2000).

Turning to the set of control variables X_t^i, two alternative models are considered: a "base" model with a lot of controls and a "small" model with a few controls. Building on Perotti (1996), Panizza (2002) and Partridge (2005), MR (2012a) insist that considering this distinction in a growth regression is important. First, the base model reduces the risk of omitted variables bias while the small model alleviates problems related to multicollinearity. Second and perhaps more importantly, the comparison of the two models allows to distinguish the indirect effect of the predictors included in the small model through those variables excluded from it. The small model includes the following controls: *human capital per capita*, the *percentage of people who live in metropolitan areas*, and the *percentage of the population above 65 years old*. Regarding the choice of the controls in the base model, MR (2012a) follow Partridge (1997). This model includes *human capital per capita*, with human capital being expected to have a positive impact on future growth rate.[53] In this paper, it is measured as the expected present value of annual earnings (discounted at 10% per year) associated with the education and experience of the population of working age (15-65 years). Notice that this measure is more accurate than the one used in MR (2012a), where a categorical variable indicating the proportion of individuals with a given educational achievement serves as a proxy for human capital. *The percentage of the population who worked on a farm* is included in order to account for the different importance of agriculture across states. The *growth rate of nonagricultural employment* during the preceding decade (e.g. non agricultural employment growth in the 1970s is used to explain per capita income growth in the 1980s) is included "in order to account for the possibility that growth in the

[52] Regional dummies included are those of Northeast, Southeast, South, and Central-West, while the North region is omitted. Time dummies included are those 1980 and 1991, while the dummy for 2000 is omitted.
[53] Human capital is measured in monetary value. Human capital per capita is considered instead of human capital in order to control for the size of each state.

previous decade could, in turn, influence growth in the following decade and be correlated with past inequality" (MR 2012a, footnote 22).[54] *Fertility* is included because it is one of the sole determinant through which income inequality impacts growth that is uncontroversial in the empirical literature: more inequality has been assessed to increase fertility rates, which in turn is detrimental to growth of per capita income (Koo and Dennis, 1999; Kremer and Chen, 2000; Perotti, 1996 ; de la Croix and Doepke, 2003). *Welfare public expenditures as a percentage of GDP* are included in order to control for the fact that, in political economy models, greater inequality implies a greater demand from the citizens for redistributive policies (including mainly welfare expenditures) that are, in turn, expected to affect negatively growth. Contrarily to MR (2012a), welfare public expenditures are decomposed into various components, namely: health and sanitation; education and culture; social security and social insurance. They are measured in percentage of GDP. Omitting welfare public expenditures or fertility variables should lead the coefficients on the inequality variables to suffer from downward bias. Lastly, the base model controls for the *nonagricultural sectoral mix* of each state. This variable is measured differently than in MR (2012a) due to the unavailability of data.[55] It consists in a categorical variable measuring the shares of nonagricultural GDP of the following sectors: industry, financial intermediation, transport and communications, public administration, and real estate activities. The omitted category contains the wholesale and retail trade sector as well as those services classified as "others" by the IBGE.[56]

Regarding the source of the fertility variable, ready-to-use fertility rates data at the state level are not available in year 1980. Consequently, they were computed directly on the original IPUM samples. The IPUMS-International provides data on the year and month of birth of the last child borne by the women respondent for the 1980 and 1991 Brazilian censuses but not for the 2000 census, unfortunately. These data would have allowed me to compute the total fertility rates, which have the advantage of not being affected by the age structure of the population (they indicate the average number of children a woman would bear

---

[54] Data on farm employment is only available for the years 1970, 1980, 1991 and 2000. Because the lag is not the same across periods, the growth rate of nonagricultural employment during the preceding decade is computed as a compounded annual growth rate.

[55] In comparison with this paper, MR (2012a) measure the importance of each sector by its share of nonagricultural employment. Moreover, their sectoral decomposition is more precise than the one in this paper because the industry sector is split into *mining*, *construction*, and *manufacturing*.

[56] These services classified as "others" by the IBGE are: hotels and restaurants, other community, social and personal service activities, private health and education, activities of private households as employers and undifferentiated production activities of private households. For more details about these classifications, see the *Classificações Nacional de Atividades Econômicas* by the *IBGE*.

if she experienced current age specific fertility rates throughout her reproductive life span). Instead of total fertility rates, I have computed general fertility rates because the data required for their computation were available in years 1980, 1991 and 2000. General fertility rates measure the number of births per 1,000 women aged 15-49. This is a proper measure of fertility but it has the disadvantage of being affected by variations in the age distribution of women within the 15-49 year old range. General fertility rates have been computed according to the following formula:

$$\left(\frac{number\ of\ births\ in\ year\ t}{number\ of\ women\ aged\ 15-49\ at\ midyear\ t}\right)*1000 \tag{9}$$

where the number of births in year t was proxied by the number of children aged strictly less than one year old in the original IPUMS samples.

2. *Results*

Following MR (2012a), the results from the pooled-OLS regression are first discussed in detail because their interpretation is straightforward. The results from the FE and RE specifications are then compared with the pooled-OLS one. The regressions including inequality measures based on both parental and non-parental circumstances (*IO64* and *IE64*) are considered as the benchmark ones. Table 5a shows the pooled-OLS results with four different ways to include these measures of inequality in equation (9), as in MR (2012a). The first one includes total inequality alone. The second one includes total inequality and IO. The third one includes total inequality and the share of IO in proportion of total inequality (named IOR for "relative IO"). This third specification has the advantage over the second one of controlling for the fact that IO is a component of total inequality and that both are consequently highly correlated, which might affect the estimation results. However, it has the disadvantage that IOR may vary even when IO does not (since it is the ratio of IO to total inequality). Lastly, the fourth specification includes both IO and IE. This is the preferred specification of MR (2012a) because it isolates the effect of each component of total inequality. We can see that inequality measures are never significant, whatever the specification considered. The signs on total inequality are positive, and those on IO vary across the specifications, which shows that the effect of IO is not robust. The coefficients on IO are always negative in the small model but always positive in the base model. This suggests that that the indirect effect of inequality through fertility and welfare expenditures concerns mainly IO instead of total income inequality. Interestingly, the coefficient on IE is negative whereas the "inequality as cholesterol hypothesis" predicts the contrary. Notice also that results on the control variables make sense and that they remain relatively stable across specifications. Mainly, the coefficient on initial GDP per capita is negative (confirming the hypothesis of a convergence across states) but looses its significance when all the controls from the base model are included. Human capital is always positive and significant, as expected. Interestingly, welfare public expenditures have no effect on growth, except those in education, which seem to be beneficial.

Table 5a: Pooled OLS on total inequality, IO64 and IE64

VARIABLES	small itot	base itot	small itot io64	base itot io64	small itot io64r	base itot io64r	small ie64 io64	base ie64 io64
Total inequality	0.66	1.09	0.97	0.84	0.76	1.03		
	(1.152)	(0.955)	(1.202)	(1.446)	(1.136)	(1.096)		
IO64			-0.76	0.55			-0.39	1.23
			(1.251)	(2.029)			(1.021)	(1.634)
IO64R					-0.23	0.11		
					(0.384)	(0.655)		
IE64							-0.25	-0.67
							(1.613)	(1.678)
Human capital	0.04**	0.05*	0.04**	0.05*	0.04**	0.05*	0.05**	0.05*
	(0.021)	(0.025)	(0.020)	(0.025)	(0.020)	(0.025)	(0.022)	(0.025)
Proportion urban	-0.14		-0.08		-0.08		-0.16	
	(0.650)		(0.675)		(0.677)		(0.717)	
Proportion elderly	-6.25*		-6.77*		-6.72*		-6.27	
	(3.495)		(3.670)		(3.654)		(3.737)	
GDP per capita	-0.09***	-0.03	-0.09***	-0.03	-0.09***	-0.03	-0.09***	-0.04
	(0.028)	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)	(0.030)	(0.029)
Proportion rural employment		1.43*		1.42*		1.42*		1.49**
		(0.698)		(0.710)		(0.713)		(0.695)
Lag rural empl. growth		-3.62**		-3.68**		-3.65**		-4.13**
		(1.431)		(1.462)		(1.453)		(1.488)
Fertility		-0.00		-0.00		-0.00		-0.00
		(0.002)		(0.002)		(0.003)		(0.002)
Welfare exp. (Health & Sanit.)		-0.00		-0.00		-0.00		-0.00
		(0.003)		(0.003)		(0.003)		(0.003)
Welfare exp. (Educ. & Culture)		0.01*		0.01*		0.01*		0.01*
		(0.004)		(0.004)		(0.004)		(0.004)
Welfare exp. (Social insurance)		-0.00		0.00		-0.00		0.00
		(0.003)		(0.003)		(0.003)		(0.003)
Sectoral mix (Industry)		1.00*		1.05*		1.03*		1.17*
		(0.550)		(0.594)		(0.603)		(0.625)
Sectoral mix (Financial int.)		0.60		0.67		0.65		0.97
		(0.838)		(0.944)		(0.959)		(0.874)
Sectoral mix (Transport & com.)		3.78*		3.88**		3.86**		4.07**
		(1.870)		(1.745)		(1.737)		(1.737)
Sectoral mix (Public admin.)		0.34		0.40		0.39		0.61
		(1.109)		(1.198)		(1.226)		(1.157)
Sectoral mix (Real estate)		-2.36**		-2.25*		-2.28*		-1.92*
		(1.093)		(1.227)		(1.258)		(1.046)

Table 5a (continued):

VARIABLES	small itot	base itot	small itot io64	base itot io64	small itot io64r	base itot io64r	small ie64 io64	base ie64 io64
Dummy 1980	0.19	0.36	0.18	0.38	0.18	0.37	0.29	0.52
	(0.275)	(0.356)	(0.269)	(0.345)	(0.268)	(0.345)	(0.280)	(0.355)
Dummy 1991	-0.19	0.03	-0.21	0.05	-0.21	0.04	-0.09	0.18
	(0.208)	(0.223)	(0.200)	(0.218)	(0.200)	(0.219)	(0.208)	(0.232)
Dummy Northeast	-0.10	-0.18	-0.11	-0.17	-0.11	-0.17	-0.07	-0.15
	(0.140)	(0.129)	(0.144)	(0.135)	(0.143)	(0.136)	(0.157)	(0.142)
Dummy Southeast	-0.04	-0.24	-0.05	-0.23	-0.05	-0.23	-0.04	-0.24
	(0.127)	(0.233)	(0.131)	(0.245)	(0.131)	(0.246)	(0.138)	(0.247)
Dummy South	0.01	-0.31	0.01	-0.30	0.01	-0.30	0.01	-0.31
	(0.143)	(0.264)	(0.145)	(0.281)	(0.145)	(0.283)	(0.147)	(0.281)
Dummy Central-West	-0.14	-0.06	-0.15	-0.06	-0.16	-0.06	-0.12	-0.06
	(0.116)	(0.198)	(0.128)	(0.201)	(0.128)	(0.202)	(0.132)	(0.201)
Constant	-0.11	-1.80**	-0.14	-1.85**	-0.07	-1.86**	0.03	-1.76*
	(0.385)	(0.848)	(0.409)	(0.840)	(0.365)	(0.874)	(0.387)	(0.863)
Observations	80	76	80	76	80	76	80	76
R-squared	0.481	0.695	0.485	0.695	0.485	0.695	0.480	0.694

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5b shows the results from the pooled-OLS regressions when inequality measures based on non-parental circumstances only (*IO4* and *IE4*) are included. This serves as a robustness check: i) against the possibility that the imputation procedure considered in section III has not worked properly; ii) against the possibility that the effect of IO on growth depends on the circumstances considered. Results are similar in the sense that inequality measures are never significant (but their sign is now always positive). Because the sign and significance of the control variables remain stable across specifications, only the results on the inequality measures are shown in table 5b and in the next tables of this paper for expositional simplicity reasons (and as MR (2012a) in their sensitivity analysis).

Table 5b: Pooled OLS on total inequality, IO4 and IE4

VARIABLES	small itot io4	base itot io4	small itot io4r	base itot io4r	small itot io4r	base itot io4r
Total inequality	0.27	0.77	0.34	0.88		
	(1.551)	(1.169)	(1.266)	(1.007)		
IO4	2.75	2.92			3.02	3.69
	(7.863)	(4.893)			(6.930)	(4.273)
IO4R			1.44	1.60		
			(2.607)	(1.768)		
IE4					0.27	0.77
					(1.551)	(1.169)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Tables 6a and 6b display the results of the same regressions considered so far when initial GDP per capita is not included as a control variable (that is, when Brazilian states are not anymore interpreted as being in their steady-state equilibrium-path). Results are highly similar. The only difference is that the coefficients on inequality measures tend to increase. Additionally, the coefficient on initial human capital per capita (not shown) decreases such that it becomes insignificant now (and even negative in some specifications). This is not surprising because it is now this variable that captures the convergence effect across states.

Table 6a: Pooled OLS on total inequality, IO64 and IE64 without initial GDP

VARIABLES	small itot	base itot	small itot io64	base itot io64	small itot io64r	base itot io64r	small itot io64r	base itot io64r
Total inequality	1.22	1.29	1.51	1.13	1.31	1.26		
	(1.162)	(0.855)	(1.284)	(1.264)	(1.165)	(0.957)		
IO64			-0.70	0.37			-0.13	0.91
			(1.498)	(1.927)			(1.220)	(1.564)
IO64R					-0.21	0.07		
					(0.457)	(0.622)		
IE64							-0.31	0.24
							(1.809)	(1.409)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6b: Pooled OLS on total inequality, IO4 and IE4 without initial GDP

VARIABLES	small itot io4	base itot io4	small itot io4r	base itot io4r	small itot io4r	base itot io4r
Total inequality	1.14	1.07	1.06	1.14		
	(1.185)	(1.062)	(1.028)	(0.911)		
IO4	0.55	2.17			1.69	3.24
	(7.724)	(4.685)			(7.251)	(4.114)
IO4R			0.74	1.33		
			(2.655)	(1.691)		
IE4					1.14	1.07
					(1.185)	(1.062)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Tables 7a and 7b show the results of the FE and RE regressions on the preferred specification: when both IO, IE and initial GDP are included (as in MR 2012a). There is not much things to say apart that results remain similar: the effect of all inequality measures is neither robust nor significant across specifications.

Table 7a: RE and FE on total inequality, IO64 and IE64

VARIABLES	RE small	RE base	FE small	FE base
IO64	-0.25	2.17	0.09	-2.03
	(0.929)	(1.540)	(1.345)	(1.465)
IE64	-1.29	-0.97	0.49	2.60
	(1.216)	(1.611)	(2.090)	(1.706)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 7b: RE and FE on total inequality, IO4 and IE4

VARIABLES	RE small	RE base	FE small	FE base
IO4	0.05	3.74	7.90	-0.25
	(5.062)	(4.907)	(11.588)	(5.961)
IE4	-0.45	1.20	-0.18	1.63
	(1.481)	(1.247)	(2.836)	(1.521)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Let me point out that almost all imaginable specifications have been tested even though only the main ones from M&R (2012a) are displayed here for sake of comparability and of expositional simplicity. Among others, the following specifications have been tried: when IO or IE are considered as the sole inequality measures; when the log of initial GDP per capita is considered instead of its level; when the log-linear form of equation (9) is considered; when regional and/or time fixed-effects are not considered; when a set of control variables different from MR (2012a) is considered. All convey the same message: IO, IE and total inequality do not have a significant and robust effect on growth.[57]

Lastly, I must acknowledge that the econometric specifications displayed here are suffering from numerous caveats. The main caveats are the following ones: the few number of year-state observations; the choice of the control variables that is open to criticism; and the endogeneity of inequality to growth, which is not properly addressed here. However, these caveats do not invalidate the conclusion from this empirical section. Indeed, MR (2012a) implement econometric specifications that are highly comparable with those considered here and that, consequently, are subject to the exactly the same caveats. Yet, MR (2012a) find that in US states – contrarily to the present paper in Brazilian states – IO and IE have a robust and significant impact on growth (negative for IO, positive for IE) while total income inequality has a positive but not robustly significant impact on growth.

[57] IO is found to be significant but with the wrong sign (ie. positive) in some of the specifications where the log of initial GDP per capita is considered instead of its level.

V. Conclusion

The concept of inequality of opportunity (IO) is relatively novel in the economic literature. A lot remains to be said about its measurement and about its instrumental effect on some key economic variables such as growth. This paper investigates each of these two issues in turn because one has to think about the meaning of IO *per se* if one is to understand how IO might impact growth. It first tries to reconcile two of the most prominent approaches to the measurement of IO in the recent literature. The first approach is from Checchi and Peragine (2010), who define IO as the inequality between groups of people sharing the same circumstances – those factors affecting income that are morally unfair in the sense that individuals have no control over them (eg. gender or skin color) – and relies on a standard decomposition of inequality into its within and between components in order to estimate IO non-parametrically. The second approach is from Ferreira and Gignoux (2011), building on Bourguignon et al. (2007a), who devise a parametric estimate of IO. This approach has the main advantage of allowing to represent individual income as a function of circumstances and efforts and to distinguish conceptually the direct effect of circumstances on income from their indirect effect through efforts. However, it has the disadvantage of yielding a measure of IO which is only an approximation of the one from Checchi and Peragine (2010). The synthesis proposed in this paper allows to retain only the desirable properties of these two approaches. This synthesis helps to interpret more clearly what the abstract concepts of circumstances, efforts and *equality* of opportunity refer to.

Turning to the instrumental effect of IO on growth, this paper provides an independent replication of Marrero and Rodríguez (2012a), the first and only study having investigated this relationship empirically so far. The two authors find some robust evidences that IO has a significantly negative impact on growth in US states. The present study shows that these findings cannot be generalized to the case of Brazilian states: the effect of IO is never found to be significantly different from zero, whatever the econometric specification used (pooled-OLS, random or fixed effects). This no-result is both important and disappointing in the sense that the "inequality as cholesterol hypothesis" – total income inequality would encompass a good sort of inequality (the inequality of effort) and bad one (IO) – validated by Marrero and Rodríguez' study was offering a convincing explanation for the absence of consensus in the growth-inequality literature: the effect of total inequality would depend on the sort of inequality that dominates.

The fact that IO has no effect on growth in Brazilian states but that it has a strong detrimental effect in US states is in line with Barro (2000), who posits that the effect of total income inequality depends on the level of development of the country considered. Studying empirically whether the results from Barro (2000) hold true regarding IO instead of total income inequality is an area for future research. An explanation for the no-result found in this paper might be that the effect of IO on growth is more long-term than the one that has been considered here (nine years). In any case, this no-result is surprising and I must acknowledge my troubles for explaining why IO would not have a beneficial effect on economic performance: *equality* of opportunity, after all, means that no one in a society is hampered from expressing his full talent and potential, which should prompt human capital accumulation and entrepreneurial dynamism. More simply, perhaps the explanation for the absence of a robust effect of IO across countries is that what is fair is not always efficient: IO might be a more appropriate concept than income inequality for the purpose of assessing whether an income distribution is fair, not whether an income distribution is efficient.

References

Alesina, A., et D. Rodrik. 1994. « Distributive Politics and Economic Growth ». *The Quarterly Journal of Economics* 109 (2): 465-490. doi:10.2307/2118470.

Alesina, Alberto, et Roberto Perotti. 1996. « Income distribution, political instability, and investment ». *European Economic Review* 40 (6): 1203-1228. doi:10.1016/0014-2921(95)00030-5.

Allison, P. D. 2001. *Missing data* (No. 136). SAGE Publications, Incorporated.

Arneson, Richard J. 1989. « Equality and Equal Opportunity for Welfare ». *Philosophical Studies* 56 (1): 77–93.

Baraldi, A. N., & Enders, C. K. 2010. « An introduction to modern missing data analyses ». *Journal of School Psychology*, *48*(1), 5-37.

Barro, Robert J. 2000. « Inequality and Growth in a Panel of Countries ». *Journal of Economic Growth* 5 (1) (mars 1): 5-32. doi:10.1023/A:1009850119329.

Benabou, Roland. 1996. « Inequality and Growth ». *NBER Macroeconomics Annual* 11: 11. doi:10.2307/3585187.

Bertola, Giuseppe. 1993. « Factor Shares and Savings in Endogenous Growth ». *American Economic Review* 83 (5): 1184-98.

Birdsall, Nancy, David Ross, et Richard Sabot. 1995. « Inequality and Growth Reconsidered: Lessons from East Asia ». *The World Bank Economic Review* 9 (3): 477-508. doi:10.1093/wber/9.3.477.

Björklund, A., Jäntti, M., & Roemer, J. E. (2011). *Equality of opportunity and the distribution of long-run income in Sweden* (No. 5466). Discussion paper series//Forschungsinstitut zur Zukunft der Arbeit.

Bleaney, Michael, et Akira Nishiyama. 2004. « Income inequality and growth—does the relationship vary with the income level? » *Economics Letters* 84 (3): 349-355. doi:10.1016/j.econlet.2004.03.004.

Bourguignon, F. 1979. « Decomposable income inequality measures ». *Econometrica: Journal of the Econometric Society*: 901–920.

———. (1993) « Croissance, distribution et ressources humaines: comparaison internationale et spécificités régionales. » *Revue d'Economie du Développement* 4 (December): 3-35.

Bourguignon, François, Francisco H. G. Ferreira, and Michael Walton. 2007b. « Equity, Efficiency and Inequality Traps: A Research Agenda » *Journal of Economic Inequality*, 5(2), 235–56

Brunori, P., Ferreira, F., & Peragine, V. 2013. « Inequality of opportunity, income inequality and economic mobility: some international comparisons. » *World Bank Policy Research Working Paper*, (6304).

Chaudhuri, Shubham , et Martin Ravallion. 2006. « Partially Awakened Giants: Uneven Growth in China and India » *World Bank Policy Research Working Paper 4069*

Checchi, Daniele, et Vito Peragine. 2010. « Inequality of Opportunity in Italy ». *The Journal of Economic Inequality* 8 (4): 429-450. doi:10.1007/s10888-009-9118-3.

Checchi, Daniele, et Vitorocco Peragine. 2005. *Regional Disparities and Inequality of Opportunity: The Case of Italy*. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=869006.

Clarke, George R.G. 1995. « More evidence on income distribution and growth ». *Journal of Development Economics* 47 (2): 403-427. doi:10.1016/0304-3878(94)00069-O.

Cleveland, L. L., Davern, M., & Ruggles, S. 2011. « Drawing Statistical Inferences from International Census Data ».

Cogneau, D., et C. Guénard. 2002. « Can a Relation Be Found Between Inequality and Growth? » *DT DIAL* 3: 39.

Cohen, G. A. 1989. « On the currency of egalitarian justice ». *Ethics* 99 (4): 906–944.

Cowell, F. A. 1980. « On the structure of additive inequality measures ». *The Review of Economic Studies*: 521–531.

Dahan, Momi, et Daniel Tsiddon. 1998. « Demographic Transition, Income Distribution, and Economic Growth ». *Journal of Economic Growth* 3 (1) (mars 1): 29‑52. doi:10.1023/A:1009769930916.

Davern, M., & Strief, J. 2008. « IPUMS User Note: Issues Concerning the Calculation of Standard Errors (ie, variance estimation) Using IPUMS Data Products. »

De la Croix, D. and Doepke, M. (2003), "Inequality and growth: why differential fertility matters", *American Economic Review*, 93, 1091-1113.

Deininger, Klaus, et Lyn Squire. 1996. « A New Data Set Measuring Income Inequality ». *The World Bank Economic Review* 10 (3): 565‑591. doi:10.1093/wber/10.3.565.

———. 1998. « New ways of looking at old issues: inequality and growth ». *Journal of Development Economics* 57 (2): 259‑287. doi:10.1016/S0304-3878(98)00099-6.

Dworkin, R. 1981. « What is equality? Part 2: Equality of resources ». *Philosophy & Public Affairs*: 283–345.

Easterly, William. 2001. « The Middle Class Consensus and Economic Development ». *Journal of Economic Growth* 6 (4): 317‑335. doi:10.1023/A:1012786330095.

———. 2007. « Inequality does cause underdevelopment: Insights from a new instrument ». *Journal of Development Economics* 84 (2): 755‑776. doi:10.1016/j.jdeveco.2006.11.002.

Ehrhart, C. 2009. « The effects of inequality on growth: a survey of the theoretical and empirical literature ». *ECINEQ WP* 107. http://www.ecineq.org/milano/WP/Ecineq2009-107.pdf.

Figueiredo, E. A., & Ziegelmann, F. A. 2010. « Estimation of opportunity inequality in Brazil using nonparametric local logistic régression ». *The Journal of Development Studies*, *46*(9), 1593-1606.

Fleurbaey, M. and V. Peragine. 2013. "Ex Ante versus Ex Post Equality of Opportunity," *Economica* 80. 118–130 doi:10.1111/j.1468-0335.2012.00941.x.

Forbes, K. J. 2000. « A Reassessment of the Relationship between Inequality and Growth ». *American economic review*: 869–887.

Foster, J. E. 1985. « Inequality measurement ». *Fair allocation* 33: 31–68.

Foster, J. E., et A. A. Shneyerov. 2000. « Path independent inequality measures ». *Journal of Economic Theory* 91 (2): 199–222.

Ferreira, Francisco, and Jérémie Gignoux. "The measurement of inequality of opportunity: theory and an application to Latin America." *World Bank Policy Research Working Paper Series, Vol* (2008).

Galor, Oded, et Hyoungsoo Zang. 1997. « Fertility, income distribution, and economic growth: Theory and cross-country evidence ». *Japan and the World Economy* 9 (2) (mai): 197‑229. doi:10.1016/S0922-1425(96)00245-9.

Galor, Oded, et Joseph Zeira. 1993. « Income Distribution and Macroeconomics ». *The Review of Economic Studies* 60 (1) (janvier 1): 35‑52. doi:10.2307/2297811.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. 1997. « Analysis with missing data in prevention research ». In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325_366). Washington, DC: American Psychological Association.

Herzer, Dierk, et Sebastian Vollmer. 2012. « Inequality and Growth: Evidence from Panel Cointegration ». *The Journal of Economic Inequality* 10 (4): 489‑503. doi:10.1007/s10888-011-9171-6.

Instituto Brasileiro de Geografia e Estatística (IBGE). 2003. Classificação Nacional de Atividades Econômicas. Versão 1.0

Khoo, L. and Dennis, B. 1999. « Income inequality, fertility choice, and economic growth: theory and evidence ». *Harvard Institute for International Development* (*HIID*), *Development Discussion Paper 687*.

Kremer, M. and Chen, D. 2002. « Income distribution dynamics with endogenous fertility », *Journal of Economic Growth*, 7, 227-258.

Kuznets, S. 1955. « Economic growth and income inequality ». *The American Economic Review* 45 (1): 1–28.

Li, Hongyi, et Heng-fu Zou. 1998. « Income Inequality Is Not Harmful for Growth: Theory and Evidence ». *Review of Development Economics* 2 (3): 318–334. doi:10.1111/1467-9361.00045.

Little, R. J. 1995. « Modeling the drop-out mechanism in repeated-measures studies ». *Journal of the American Statistical Association*, *90*(431), 1112-1121.

Marrero, G. A., et J. G. Rodríguez. 2012a. « Inequality of Opportunity and Growth ». *Documento de Trabajo/Working Paper Serie Economía. Revised and resubmit at Journal of Development Economics*. http://www.uv.es/~estevev/DT-E-2011-05.pdf.

Marrero, G. A., et J. G. Rodríguez. 2012b. « Inequality of Opportunity and Inequality of Effort: A Canonical Growth Model ». *Paper Prepared for the 32nd General Conference of The International Association for Research in Income and Wealth*

Minnesota Population Center. 2011. *Integrated Public Use Microdata Series, International: Version 6.1* [Machine-readable database]. Minneapolis: University of Minnesota.

Newman, D. A. 2003. « Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques ». *Organizational Research Methods*, *6*(3), 328-362.

Nilsson, W. 2005. « Opportunities, preferences and incomes ». *Umea Economic Studies*, *649*.

O'Neill, D., & Sweetman, O. D. Van de gaer .2000. « Equality of opportunity and Kernel density estimation: An application to intergenerational mobility ». *Advances in Econometrics*, *14*, 259-274.

Ooghe, E., & Schokkaert, E. 2007. « Equality of opportunity versus equality of opportunity sets ». *Social Choice and Welfare*, *28*(2), 209-230.

Panizza, U. 2002. « Income inequality and economic growth: evidence from American Data », *Journal of Economic Growth*, 7, 25-41.

Partridge, M. D. 1997. « Is Inequality Harmful for Growth? Comment ». *American Economic Review*, 87, 5, 1019-1032.

———. 2005. « Does income distribution affect U.S. state economic growth? », *Journal of Regional Science,* 45, 363-394.

Peragine, V. 2000. « Ranking income distributions on the basis of equality of opportunity ». *Dipartimento di economia pubblica e territoriale, Università di Pavia*. http://www.dse.uniba.it/Seminari/rank1.PDF.

Peragine, V. 2004. « Ranking income distributions according to equality of opportunity ». *The Journal of Economic Inequality*, *2*(1), 11-30.

Perotti, Roberto. 1996. « Growth, Income Distribution, and Democracy: What the Data Say ». *Journal of Economic Growth* 1 (2): 149-187. doi:10.1007/BF00138861.

Persson, Torsten, et Guido Tabellini. 1994. « Is Inequality Harmful for Growth? » *American Economic Review* 84 (3): 600-621.

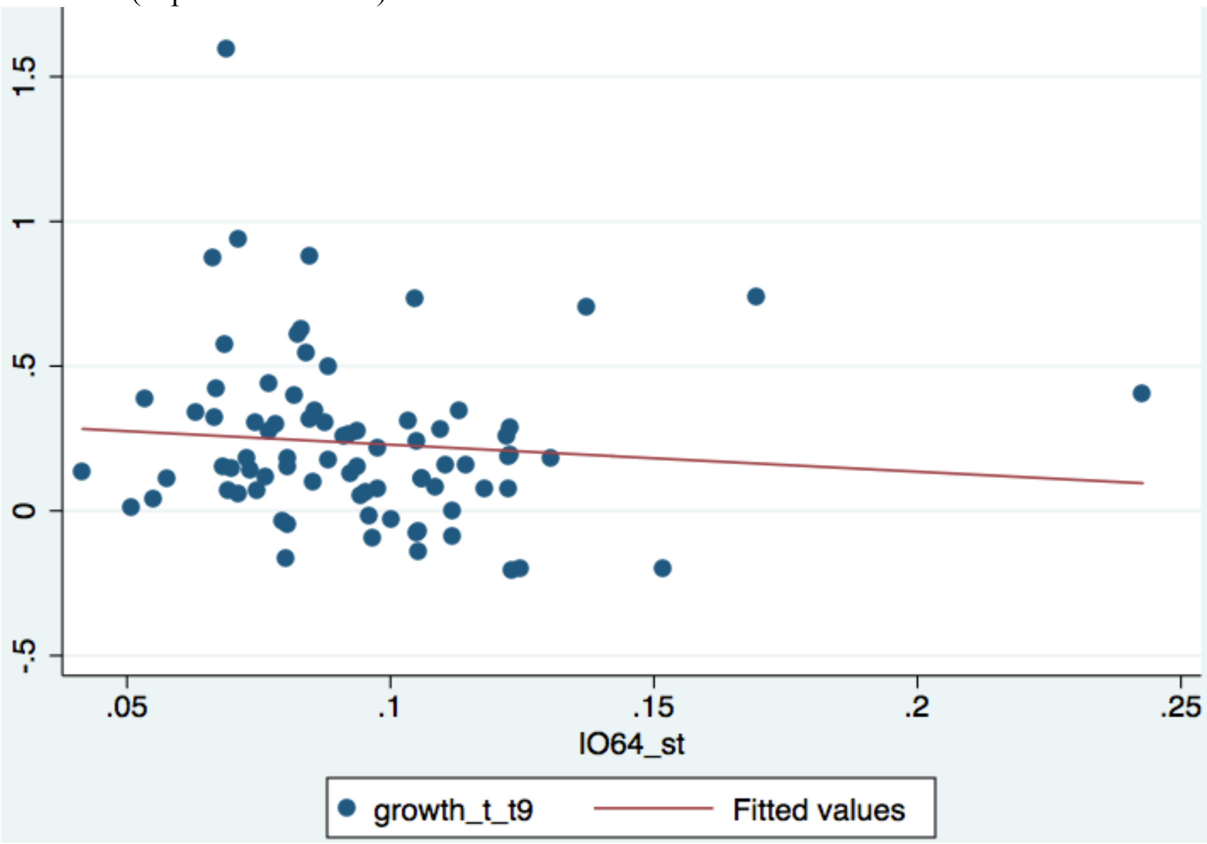
Pignataro, G. 2012. Equality of opportunity: Policy and measurement paradigms. *Journal of Economic Surveys*, *26*(5), 800-834.

Piketty, Thomas. 1997. « The Dynamics of the Wealth Distribution and the Interest Rate with Credit Rationing ». *The Review of Economic Studies* 64 (2) (janvier 4): 173-189.

doi:10.2307/2971708.
Ramos, X., & Van de Gaer, D. 2012. Empirical approaches to inequality of opportunity: Principles, measures, and evidence.
Roemer, J. E. 1993. « A pragmatic theory of responsibility for the egalitarian planner ». *Philosophy & Public Affairs* 22 (2): 146–166.
———. 1998. Equality of Opportunity, Harvard University Press, Cambridge.
Rubin, D. B. 1977. « Formalizing Subjective Notion about the Effect of Nonrespondents in Sample Surveys ». *Journal of the American Statistical Association* 72: 538-543.
———. 1987. « Multiple imputation for non-response in surveys ». New York: John Wiley & Sons.
———. 1996. « Multiple imputation after 18+ years ». *Journal of the American Statistical Association*. 91. 473 - 489.
Ruiz-Castillo, Javier. 2003. « The measurement of the inequality of opportunities ». *Research on Economic Inequality* 9: 1-34. doi:10.1016/S1049-2585(03)09002-1.
Saint-Paul, Gilles, et Thierry Verdier. 1993. « Education, democracy and growth ». *Journal of Development Economics* 42 (2) (décembre): 399-407. doi:10.1016/0304-3878(93)90027-K.
Sen Amartya, K. 1985. *Commodities and capabilities*. North Holland, Amsterdam.
Shorrocks, A. F. 1980. « The class of additively decomposable inequality measures ». *Econometrica: Journal of the Econometric Society*: 613–625.
Sobek, M., & Kennedy, S. 2009. « The development of family interrelationship variables for international census data, Minnesota Population Center. »
van de Gaer, D. 1993. *Equality of opportunity and investment in human capital*. Katholieke Universiteit Leuven, Faculteit der Economische en Toegepaste Economische Wetenschappen.
Wayman, J. C. 2003. « Multiple imputation for missing data: What is it and how can I use it ». In *Annual Meeting of the American Educational Research Association, Chicago, IL* (pp. 2-16).
World Bank. 2006. World Development Report : Equity and Development, The World Bank and Oxford University Press, Washington, 2006.
Yusuf, A. A. 2005. *A Survey on Growth and Inequality: Does Improved Inequality Data Have Anything to Say?* Department of Economics, Padjadjaran University. http://equitablepolicy.org/wpaper/200501.pdf.

Appendix

Figure 2: Simple OLS regression of *IO64* (explanatory variable) regressed on growth (explained variable)



Note: Growth of GDP per capita is on the Y-axis; IO64 on X-axis. Regression on 80 year-states observations. The regression line has a coefficient of -0.93 but is strongly insignificant (p-val= 0.477).

Table 4: Inequality in Brazilian states (1980-2000)

State	Total inequality			IO64			IO4			IE64			IE4		
	1980	**1991**	**2000**	**1980**	**1991**	**2000**	**1980**	**1991**	**2000**	**1980**	**1991**	**2000**	**1980**	**1991**	**2000**
Rondônia	.316302	.345381	.277184	.109611	.130617	.082695	.014078	.009289	.014043	.188947	.227461	.192775	.302224	.336092	.263142
Acre	.255497	.321797	.273192	.041573	.050871	.104935	.008658	.01288	.012656	.147313	.184536	.173236	.24684	.308917	.260537
Amazonas	.281967	.321385	.258874	.084216	.096758	.071156	.026722	.020126	.015165	.19103	.228539	.181362	.255245	.301259	.243708
Roraima	.298027	.299799	.285378	.242722	.105433	.08466	.034112	.01242	.023157	.162628	.182771	.143703	.263915	.287378	.262221
Pará	.295885	.345821	.26862	.074636	.124826	.122298	.026707	.015958	.014941	.213013	.269084	.197716	.269178	.329863	.253679
Amapá	.274052	.290026	.278866	.07126	.123044	.081956	.01236	.009967	.011264	.169398	.191029	.156348	.261692	.280059	.267602
Tocantins	.	.381533	.281911	.	.114359	.068859	.	.021458	.025138	.	.234249	.168293	.	.360074	.256773
Maranhão	.306737	.34302	.270639	.05515	.069264	.066345	.019894	.017411	.012627	.243736	.276459	.197629	.286844	.32561	.258012
Piauí	.336355	.371183	.280667	.070042	.068328	.068662	.017125	.016076	.015525	.264474	.269274	.216017	.31923	.355106	.265142
Ceará	.383763	.403756	.285612	.093727	.094439	.063263	.030022	.020734	.014067	.287132	.302249	.201835	.353741	.383022	.271545
Rio Grande do Norte	.344765	.360576	.266769	.103604	.076379	.078473	.03026	.020054	.017815	.261676	.276894	.194272	.314505	.340522	.248954
Paraíba	.362384	.377803	.255588	.077076	.085489	.053539	.020567	.017398	.012267	.275553	.294831	.205219	.341817	.360405	.243321
Pernambuco	.37126	.390187	.288308	.122466	.096041	.072733	.044196	.029033	.019768	.266339	.277137	.192371	.327065	.361154	.26854
Alagoas	.321961	.353065	.256684	.079534	.080475	.066603	.027303	.029727	.016399	.238877	.257644	.185163	.294658	.323337	.240285
Sergipe	.333412	.356524	.250475	.088359	.080256	.077096	.024836	.021731	.012053	.232544	.239696	.185802	.308576	.334792	.238422
Bahia	.355186	.400993	.274596	.110631	.097547	.105201	.034494	.029751	.018828	.242456	.284204	.198644	.320691	.371243	.255767
Minas Gerais	.357305	.364911	.275966	.122967	.095383	.088196	.039051	.02989	.023692	.245714	.266815	.196027	.318254	.33502	.252274
Espírito Santo	.316507	.334605	.285793	.092097	.080696	.085915	.033112	.023942	.02689	.226694	.261607	.191343	.283394	.310663	.258902
Rio de Janeiro	.336213	.379048	.287451	.112014	.108649	.092462	.053633	.044214	.032079	.22654	.267026	.200248	.28258	.334834	.255372
São Paulo	.313211	.340056	.285955	.106234	.100285	.093931	.043123	.033135	.027916	.2113	.235028	.191408	.270088	.30692	.25804
Paraná	.347698	.369109	.302984	.113108	.118138	.091237	.032368	.029475	.022367	.235913	.265393	.199941	.31533	.339635	.280618
Santa Catarina	.279381	.309032	.2713	.067226	.057571	.087595	.020213	.018028	.014824	.199053	.219557	.197319	.259169	.291004	.256476
Rio Grande do Sul	.323996	.354964	.282408	.097675	.074953	.080611	.028792	.020024	.015699	.231813	.264634	.209494	.295204	.33494	.26671
Mato Grosso do Sul	.366944	.369837	.285264	.105064	.106196	.084802	.038613	.029331	.026031	.227777	.243127	.186903	.328331	.340506	.259233
Mato Grosso	.317256	.366412	.304792	.073369	.122755	.16965	.02038	.027327	.029865	.211822	.236617	.196677	.296876	.339085	.274927
Goiás	.362272	.363269	.287478	.105346	.111856	.083151	.031458	.025632	.023666	.246977	.258111	.199604	.330814	.337637	.263812
Distrito Federal	.374283	.404852	.370625	.151871	.122616	.137444	.072961	.048289	.040829	.227073	.250832	.219489	.301322	.356563	.329796